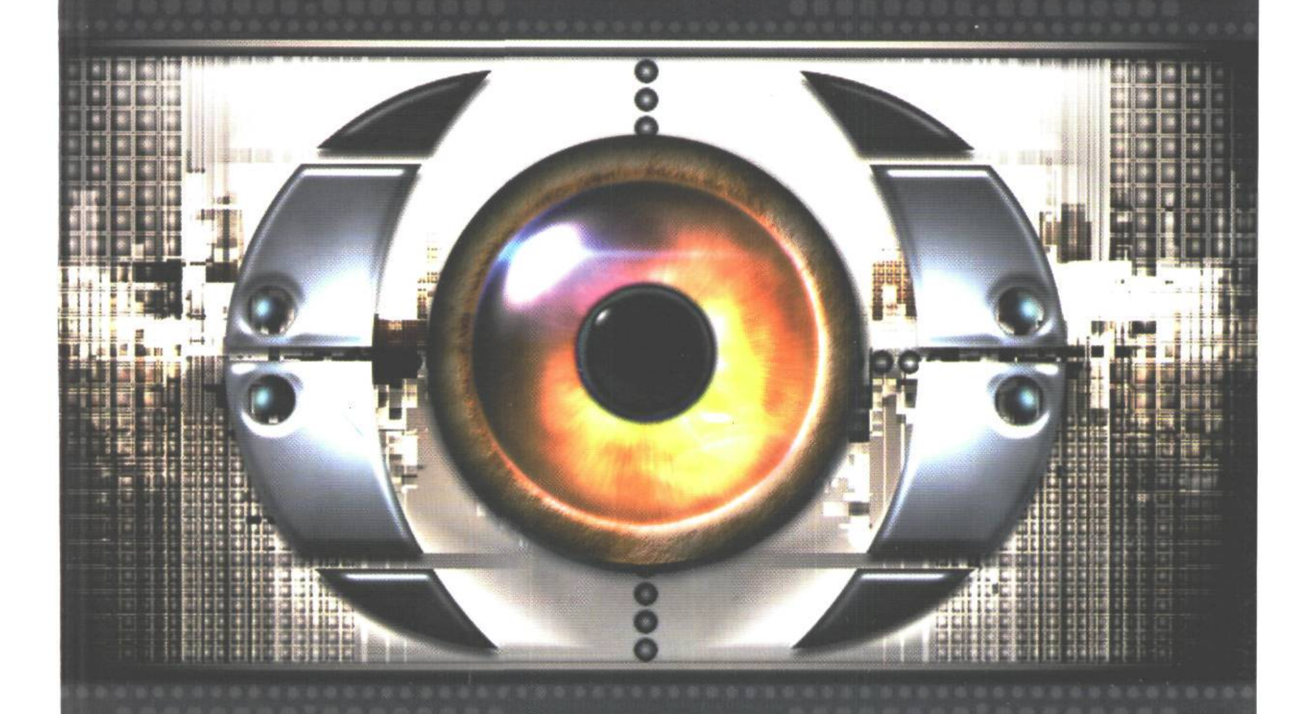
А. Наследов

# SPSS

## КОМПЬЮТЕРНЫЙ АНАЛИЗ ДАННЫХ В ПСИХОЛОГИИ И СОЦИАЛЬНЫХ НАУКАХ





А. Наследов

# SPSS

## КОМПЬЮТЕРНЫЙ АНАЛИЗ ДАННЫХ В ПСИХОЛОГИИ И СОЦИАЛЬНЫХ НАУКАХ




Москва • Санкт-Петербург • Нижний Новгород • Воронеж
Новосибирск • Ростов-на-Дону • Екатеринбург • Самара
Киев • Харьков • Минск
2005

*Андрей Дмитриевич Наследов*

# SPSS: Компьютерный анализ данных в психологии и социальных науках

| | |
|---|---|
| Главный редактор | *Е. Строганова* |
| Заведующий редакцией компьютерной литературы | *А. Кривцов* |
| Заведующий редакцией психологической литературы | *Л. Винокуров* |
| Руководитель проекта | *В. Шрага* |
| Литературный редактор | *А. Жданов* |
| Художник | *Л. Адуевская* |
| Иллюстрации | *А. Кузнецов, М. Шендерова* |
| Корректоры | *Н. Солнцева, Н. Лукина* |
| Верстка | *Р. Гришанов* |

ББК 88с11я7+32.973.23
УДК [159.9:681.3](075)

**Наследов А. Д.**

Н31 SPSS: Компьютерный анализ данных в психологии и социальных науках. — СПб.: Питер, 2005. — 416 с.: ил.

ISBN 5-318-00703-1

Эта книга является самоучителем и справочным руководством по анализу данных на компьютере с помощью программы SPSS. В ней описывается полный набор методов компьютерного анализа: от простейших способов числового, табличного и графического представления данных до самых современных и сложных многомерных. Материал книги организован таким образом, чтобы удовлетворить запросы как новичка, впервые приступающего к анализу данных на компьютере, так и опытного исследователя, желающего воспользоваться самыми современными методами. Основное содержание глав составляют пошаговые инструкции по реализации различных видов математико-статистического анализа в SPSS. Особое внимание уделяется получаемым результатам и их интерпретации. В конце книги приведен глоссарий, содержащий определения большинства статистических терминов. Издание адресовано исследователям в области психологии и социальных наук, а также широкому кругу читателей, желающих воспользоваться программой SPSS для профессионального анализа данных.



ISBN 5-318-00703-1

ООО «Питер Принт». 194044, Санкт-Петербург, пр. Б. Сампсониевский, д. 29а.
Лицензия ИД № 05784 от 07.09.01.
Налоговая льгота — общероссийский классификатор продукции ОК 005-93, том 2; 953005 — литература учебная.
Подписано в печать 29.11.04. Формат 70Х100/16. Усл. п. л. 33,54. Тираж 3000 экз. Заказ № 1251.
Отпечатано с готовых диапозитивов в ФГУП «Печатный двор» им. А. М. Горького
Министерства РФ по делам печати, телерадиовещания и средств массовых коммуникаций.
197110, Санкт-Петербург, Чкаловский пр., 15.

# Краткое содержание

# Содержание

# Предисловие

Внедрение компьютеров буквально во все сферы человеческой деятельности является на сегодняшний день, наверное, самым очевидным итогом научного прогресса. И как следовало ожидать, в существенной степени компьютеризация изменила характер самих научных исследований, в том числе в психологии и социальных науках. Компьютер обычно применяется исследователями для выполнения такой работы, которая считается самой скучной и утомительной: учет и организация исходных данных, вычисления различных показателей и пр. Это позволяет исследователю проводить более глубокий анализ данных, больше времени уделять интерпретации и выдвижению новых предположений, то есть заниматься тем, что обычно считается самым приятным и интересным в любом исследовании, но что остается за пределами возможностей компьютера.

Эта книга является учебным пособием и руководством по применению компьютерной программы SPSS — очень мощного и широко распространенного средства компьютерного анализа данных в психологии и социальных науках. SPSS — это аббревиатура от Statistical Package for the Social Science (статистический пакет для социальных наук). Как следует из названия, SPSS представляет собой множество различных программ, предназначенных для анализа данных в социальных науках. Эти программы позволяют организовывать ввод данных, гибко менять их структуру, применять к ним самые современные методы обработки или их последовательность и получать результаты в удобной и наглядной форме. Все это множество программ объединено в единую систему, обеспечивающую простой и дружественный диалог с исследователем и снабженную исчерпывающей справочной поддержкой. Благодаря такой дружественности система SPSS легко доступна для освоения даже тем, кто имеет минимальные навыки владения компьютером.

Но если SPSS настолько дружественная система, да еще обладающая справочной поддержкой, то зачем нужна эта книга? Как показывает практика, оставшись один на один со справочной системой SPSS, даже опытный исследователь, не говоря о студентах, впадает в отчаяние, безуспешно пытаясь найти в мегабайтах документации нужную информацию. К этому следует добавить языковой барьер (SPSS, как правило, англоязычная программа) и не всегда блестящее владение пользователем статистической и математической терминологией. Но даже те, кто применяет SPSS для анализа данных, испытывают трудности при необходимости реорганизации исходных данных и редактировании результатов обработки (таблиц и графиков). Именно все это и стало стимулом к написанию этой книги.

В книге содержится подробное описание большинства процедур SPSS, предназначенных для анализа данных — от простейших до самых сложных, включая различные нюансы их применения. При описании любой процедуры основными подходами являлись: максимально доступное изложение ее назначения, сведение действий по ее применению к четким пошаговым инструкциям, демонстрация ее применения на простейших примерах. Однако было бы неверно думать, что эта книга предназначена только для тех, кто хочет освоить азы компьютерного анализа данных. Она, без сомнения, будет полезна и как справочник для тех, кто уже имеет опыт работы с программой SPSS. Справочному назначению книги способствуют глоссарий и англо-русский словарь терминов, приведенные в конце книги.

Содержание первых 13 глав адресовано широкой аудитории и не требует специальных знаний в области статистики, однако оставшаяся часть книги рассчитана на людей, имеющих такие знания. В начале каждой из глав обычно кратко излагается сущность статистической процедуры, которой посвящена глава. При этом делается все возможное, чтобы исключить избыточные детали и оставить лишь ту информацию, которой достаточно для анализа данных. Говорить простым языком о сложных вещах — весьма нелегкое занятие, и пришлось приложить немало усилий, чтобы написать эту книгу. Насколько успешными были эти усилия, судить вам, читатель.

Программа SPSS снабжена подробной справочной системой. На приблизительно 3000 страницах можно найти практически любую информацию по работе с SPSS, которая обычно необходима даже опытному пользователю. Разумеется, такой объем информации немыслимо было разместить на нескольких сотнях страниц этой книги. Однако подавляющее большинство методов нашло свое отражение в этой книге, хоть временами и поверхностное, но, хочется верить, достаточное для практического применения.

# От издательства

Ваши замечания, предложения, вопросы отправляйте по адресу электронной почты comp@piter.com (издательство «Питер», компьютерная редакция).

Мы будем рады узнать ваше мнение!

Все файлы данных, упомянутые в книге, вы сможете найти по адресу http://www.piter.com/download.

Подробную информацию о наших книгах вы найдете на веб-сайте издательства http://www.piter.com.

# 1 Введение



Основное содержание этой книги составляют пошаговые инструкции по применению программы SPSS для самого разностороннего анализа данных. Книга организована так, чтобы обеспечить как новичка, так и опытного пользователя исчерпывающей информацией для самостоятельной обработки данных.

В данной главе представлено краткое содержание книги, описана ее структура, даны некоторые рекомендации по ее изучению. Если вы не знакомы с программой SPSS, то вам рекомендуется прочитать эту главу полностью, поскольку она поможет вам лучше ориентироваться в материале книги.

Если вы уже имеете опыт обработки данных при помощи программы SPSS for Windows и вас интересует лишь описание конкретной статистической процедуры, вам достаточно найти это описание в одной из глав 6-26. Эти главы не зависят друг от друга: открыв любую из них и следуя инструкциям, вы получите результат и сможете в нем разобраться. Однако даже если вы опытный пользователь, вам рекомендуется все же обратить внимание на главы 2–5, которые содержат полезные рекомендации по таким важным вопросам, как управление исходными данными, редактирование отчетов и таблиц результатов, создание и модификация графиков и диаграмм.

# Обработка данных на компьютере

Анализ данных с применением компьютера включает выполнение ряда необходимых шагов.

1. Определение структуры данных.
2. Ввод данных в компьютер в соответствии с их структурой и требованиями программы.
3. Задание метода обработки данных в соответствии с задачами исследования.

4. Получение результата обработки данных.
5. Интерпретация результата обработки.

Шаги 1 (подготовительный) и 5 (заключительный) не способна выполнить ни одна компьютерная программа — их исследователь делает сам. Помощь компьютера (шаги 2-4) заключается, в конечном итоге, в переходе от длинной последовательности чисел к более компактной их последовательности. На «вход» компьютера исследователь подает массив исходных данных, который недоступен осмыслению, но пригоден для компьютерной обработки (шаг 2). Затем исследователь дает программе команду на обработку данных в соответствии с задачей исследования и структурой данных (шаг 3). На «выходе» исследователь получает результат обработки (шаг 4) — тоже массив данных, только меньший, доступный осмыслению и содержательной интерпретации. При этом исчерпывающий анализ данных обычно требует многократной их обработки с применением разных методов.

В этой книге очень подробно и последовательно рассматривается все, что связано с компьютерной обработкой данных (шаги 2-4), кратко излагаются математические основы каждого метода, а интерпретация результатов дается на конкретных примерах.

# Необходимые знания

Эта книга станет для вас действительно эффективным руководством по компьютерной обработке данных, если вы имеете хотя бы самые общие знания о статистике и обладаете элементарными навыками работы с компьютером.

## Статистика

В начале каждой главы приводятся основные сведения, касающиеся той или иной статистической процедуры. Кроме того, в конце книги приведен глоссарий, содержащий определения большинства статистических терминов. Однако то и другое представляет собой информацию справочного характера и не рассчитаны на новичков. Несмотря на то что практически любой человек с помощью этой книги может работать с пакетом SPSS и получать результаты, отсутствие базовых статистических знаний не позволит ему самостоятельно решать даже простые задачи. Поэтому вы должны быть знакомы с основами статистики или, как минимум, находиться в процессе их изучения.

От главы к главе требования к знаниям статистики постепенно растут. А при описании сложных методов анализа данных (главы 15-26) даже подготовленному читателю, вероятно, потребуется обращение к специальной литературе[1].

---

[1] Математико-статистические основы большинства описываемых методов читатель может найти в книге: *Наследов А. Д.* Математические методы психологического исследования. Анализ и интерпретация данных. — СПб.: Речь, 2004.

### Компьютер

Требования к знанию компьютера весьма скромные: вам необходимо лишь уметь включать компьютер, а также пользоваться клавиатурой и мышью.

На вашем компьютере должны быть установлены операционная система Windows и программа SPSS for Windows (желательно версии от 9.0 и выше), а значок программы — находиться на рабочем столе.

## Версии

По всем параметрам SPSS for Windows является сложным и мощным статистическим пакетом. Однако несмотря на сложность, средства взаимодействия входящих в пакет программ с пользователем являются весьма дружественными. С помощью пакета SPSS можно проводить практически любой анализ данных, известный в социальных науках, а последние версии программы находят применение в самых разных научных областях и в мире бизнеса.

Основой для данной книги послужили версии 11 и 11.5 пакета SPSS for Windows. Говоря точнее, снимки экрана, присутствующие в книге, соответствуют версиям 11 и 11.5. Однако почти весь изложенный материал может быть с успехом применен и к более ранним версиям, таким как SPSS 7.0-7.5, SPSS 8.0–11.5. Даже для версий 6.0-6.2 большая часть приведенных процедур окажется вполне актуальной; основные отличия будут иметь место в интерфейсе программ: названиях диалоговых окон, их виде и т. п. В целом можно сказать, что соответствие данной книги версии 11.0 составляет 100 %, версиям 9.0 и 10.0 — 99 %, версиям 7.0-8.0 - 95 %, версиям 6.0-6.2 - 90 %.

В книге речь пойдет о статистических процедурах, реализованных в трех модулях и наиболее часто используемых в социальных исследованиях. В контексте SPSS *модуль* — это набор определенных статистических операций. Мы опишем операции, относящиеся к основному системному модулю (base system unit), модулю дополнительных моделей (advanced models unit) и модулю регрессионных моделей (regression models unit). Вместе с модулями фирма SPSS создала подробную справочную систему, содержащую информацию обо всех процедурах, поддерживаемых этими модулями. Однако для начинающего или недостаточно опытного пользователя справочная система слишком сложна, а ее объем, составляющий около 3000 печатных страниц, затрудняет поиск нужной информации. В определенной степени это и послужило поводом к написанию книги, в которой в компактной форме представлено практически все, что может оказаться полезным на практике.

## Содержание

Книга содержит описания большинства методов обработки данных. Решение задачи анализа данных, как отмечалось, происходит в три основных этапа:

1. Ввод данных в компьютер и их структурирование для дальнейшей обработки программой.

2. Определение типа анализа, который необходим для решения задачи.
3. Получение и расшифровка результатов обработки исходных данных.

В главе 2 пойдет речь об основах работы с пакетом SPSS: типах окон, использовании панелей инструментов и меню, просмотре, редактировании, сохранении и печати результатов обработки и т. д. Несмотря на то что глава в основном рассчитана на начинающих, в ней содержится немало информации, полезной и более опытным пользователям. Глава 3 посвящена первому из трех описанных выше этапов и включает темы создания, редактирования и форматирования файлов данных. Редактор данных SPSS является инструментом, позволяющим без труда создавать файлы данных и управлять их структурой.

В главах 4 и 5 рассматриваются такие важные вопросы, как модификация и преобразование данных, а также создание диаграмм. Глава 4 описывает различные действия с данными: создание новых переменных, сортировка, изменение структуры, объединение файлов, выбор подмножества данных для анализа. В главе 5 рассмотрены основные процедуры работы с диаграммами; заметим, однако, что большая часть материала, касающаяся диаграмм, излагается в других главах книги по мере необходимости.

Все главы начиная с главы 6 посвящены двум оставшимся этапам решения задачи анализа данных. В них идет речь о различных методах обработки данных и о том, каким образом интерпретировать их результаты. Каждая из глав 6-25 не зависит от других; это означает, что если вы захотите выполнить какой-либо вариант анализа данных, даже не обладая начальными знаниями, то сможете, открыв соответствующую главу и следуя представленным инструкциям, получить результат. При необходимости в инструкциях приводятся ссылки на другие главы.

Как было отмечено ранее, книга охватывает содержимое трех модулей SPSS: Base System, Advanced Models и Regression Models. Поскольку на практике иногда производится неполная установка пакета SPSS, некоторые из модулей могут оказаться недоступными (кроме модуля Base System, который устанавливается всегда). Для выполнения подавляющего большинства методов анализа достаточно модуля Base System. Модули Advanced Models и Regression Models позволяют дополнительно выполнять следующие методы анализа:

- многомерный дисперсионный анализ (MANOVA) и многомерный ковариационный анализ (MANCOVA);
- многомерный дисперсионный анализ с повторными измерениями (repeated-measures MANOVA);
- логистическая регрессия (logistic regression);
- логлинейный анализ (loglinear analysis).

В конце книги находятся глоссарий и англо-русский словарь терминов, к которым вам, вероятно, придется обращаться особенно часто. Глоссарий содержит краткие определения основных статистических терминов, а наличие англо-русского словаря обусловлено тем, что вам, скорее всего, придется пользоваться англоязычной версией SPSS.

# Файлы примеров

В большинстве глав книги при помощи разных методов анализируется один и тот же файл данных. Для последних глав были разработаны отдельные примеры, полнее демонстрирующие применение специфических возможностей пакета SPSS. Все файлы доступны на сайте издательства (http://www.piter.com/download). Наличие готовых исходных файлов (включая первый) полностью избавит вас от необходимости набирать данные для них вручную. Выполняя различные процедуры, вы сможете сравнивать получаемые вами результаты с «эталонными», приведенными в этой книге.

Первый файл (ex01.sav) создан таким образом, чтобы с его помощью можно было без труда продемонстрировать функционирование основных методов обработки. Он содержит информацию о 100 школьниках трех выпускных классов. О каждом из школьников имеются следующие данные:

- № (идентификационный номер);
- пол;
- класс;
- вуз (выбранный профиль вуза);
- хобби (систематическое внешкольное занятие);
- тест 1;
- тест 2;
- тест 3;
- тест 4;
- тест 5;
- средняя годовая отметка за 10-й класс;
- средняя годовая отметка за 11-й класс.

Файл ex01.sav нам потребуется, начиная уже с вводных глав 2-5. Его содержимое частично приведено в конце главы 3. При желании вы можете вручную создать собственный файл и успешно использовать его в процессе изучения почти всей первой половины книги. Разумеется, все данные уже содержатся

в готовом файле на сайте, однако вам стоит самостоятельно попрактиковаться во вводе, форматировании и других простых, но необходимых действиях с данными. Еще более полезным занятием была бы работа с собственными файлами данных.

Всего с упомянутого сайта издательства вы можете загрузить 16 файлов примеров. Из них 15 файлов содержат данные, а один файл (cross.spo) предназначен для демонстрации результатов обработки (окна вывода). В табл. 1.1 приведен список файлов примеров с указанием глав книги, в которых они используются. В указанных главах вы можете найти либо подробные описания соответствующих примеров, либо ссылки на места в книге, содержащие эти описания.

**Таблица 1.1.** Файлы примеров и главы, в которых они используются

| Файлы примеров | Главы книги |
|---|---|
| exOl.sav | 2–14 |
| cross.spo | 2 |
| ex01a.sav, ex01b.sav | 4 |
| ex02.sav | 15–16 |
| exam.sav | 17, 26 |
| help.sav | 18 |
| TestA.sav | 19 |
| TestIQ.sav | 20, 22 |
| mds1.sav, mds2.sav, mds3.sav | 21 |
| cars.sav | 22 |
| class.sav | 23 |
| helpLR.sav | 24 |
| helpLLM.sav | 25, 26 |

В каждой из глав, начиная с главы 6 и заканчивая главой 26, имеются пошаговые процедуры, в которых шаг 3 относится к инструкциям, связанным с открытием конкретного файла данных. При этом предполагается, что эти файлы данных сохранены в отдельной папке на жестком диске вашего компьютера.

# Структура глав и элементы описания

Главы 2-5, являющиеся вводными, не имеют единой структуры, поскольку едва ли возможно систематизировать начальную информацию о работе с программой. Материал остальных глав излагается в соответствии с общей структурой книги, которая выглядит следующим образом:

- Вводный раздел содержит общее описание рассматриваемой процедуры. В зависимости от сложности анализа описание занимает от 1 до 7 страниц.

- Далее идет раздел пошаговых процедур, требующихся для выполнения одной или нескольких разновидностей статистического анализа.
- Раздел печати результатов и выхода из программы посвящен, очевидно, печати полученных данных и завершающим операциям.
- В последнем разделе каждой главы представлены результаты анализа (как правило, в сокращенном виде). В этом разделе вводится необходимая терминология, позволяющая понять смысл результатов.

## Снимки экрана

Каждая глава содержит снимки экрана компьютера с фрагментами выводимых данных, элементами окон или целыми окнами, появляющимися на экране компьютера в процессе работы. На рис. 1.1 в качестве примера показано диалоговое окно Frequencies (Частоты), полученное в ходе выполнения шага 4 пошаговой процедуры из главы 6.

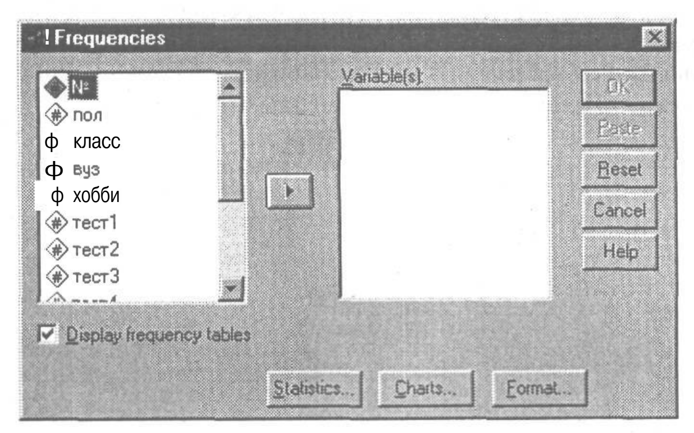



**Рис. 1.1.** Диалоговое окно Frequencies

Иногда в книге вы можете встретить рисунок, на котором будет представлено не целое окно программы, а его фрагмент, например такой, как кнопка рядом с этим абзацем или показанная ниже строка меню. Подобный рисунок призван напомнить вам, как выглядит упомянутый в текущем абзаце элемент интерфейса или описываемый фрагмент окна.




**Помимо** окон и элементов интерфейса программы в разделах «Представление результатов» вам будут встречаться большие фрагменты данных, генерируемые программой. Все такие фрагменты в каждом случае будут сопровождаться подробным описанием.

## Пошаговые процедуры

Основные статистические операции оформлены в виде пошаговых процедур, чтобы вы точно знали, что нужно сделать для решения задачи. Например, ниже представлен шаг 3 пошаговой процедуры из главы 6.

**Шаг 3** Откройте файл данных, с которым вы намерены работать (в нашем случае это файл ex01.sav). Если он расположен в текущей папке, то выполните следующие действия.

1. Выберите команду File ▸ Open ▸ Data (Файл ▸ Открытие ▸ Данные) или щелкните на кнопке Open File (Открытие файла) панели инструментов.
2. В открывшемся диалоговом окне дважды щелкните на имени ex01.sav или введите его с клавиатуры и щелкните на кнопке OK.

Как можно видеть по приведенному фрагменту пошаговой процедуры, в книге используются некоторые элементы шрифтового выделения. Так, запись File ▸ Open ▸ Data (Файл ▸ Открытие ▸ Данные) означает, что в строке меню нужно щелкнуть мышью на пункте File (Файл), в открывшемся меню раскрыть (щелчком мыши) подменю Open (Открытие) и уже в этом подменю выбрать (снова щелчком мыши) команду Data (Данные). Как вы наверно успели заметить, особым шрифтом набраны имена элементов интерфейса, файлов и переменных.

# 2 Общий обзор SPSS



Как мы упоминали в главе 1, необходимые начальные знания пользователя ограничиваются умением включать компьютер и представлением о рабочем столе Windows. В этой главе описаны все действия, позволяющие работать с пакетом SPSS for Windows. Вы узнаете, как запускать программу SPSS, как переключаться между программами с помощью панели задач и т. д. В результате у вас появятся начальные навыки работы с программой SPSS.

Если вы свободно владеете компьютером, то большую часть материала этой главы можно пропустить. Тем не менее каждому читателю вне зависимости от его опыта работы па ПК следует просмотреть эту главу «по диагонали», поскольку в ней содержится специфическая информация, касающаяся функционирования SPSS.

## Запуск программы

Пожалуй, основное, что вам необходимо (по крайней мере в рамках темы этой книги), — знать, как запустить программу SPSS. Для этого па большинстве компьютеров вам следует, щелкнув на кнопке Пуск (Start), последовательно переместить указатель мыши сначала на пункт Программы (Programs), а затем на пункт SPSS for Windows. Когда последний окажется выделенным, раскроется программная группа пакета. Щелкните мышью па команде SPSS 11.5 for Windows. После этого программа SPSS запустится. Если значок SPSS находится на рабочем столе, для запуска программы достаточно дважды щелкнуть на этом значке.

Помимо запуска SPSS вы также должны уметь переключаться между окнами программ с помощью напели задач. Для работы с пакетом SPSS это весьма актуально, поскольку он представляет собой набор из нескольких одновременно выполняющихся программ, имеющих собственные окна. При запуске SPSS вы видите

на экране единственное окно редактора данных, однако как только вы займетесь статистическим анализом ваших данных, на экране также появится *окно вывода*. В зависимости от выполняемых вами действий число открытых окон может меняться, однако окна редактора данных и вывода, как правило, присутствуют на экране постоянно.

Иногда пакет SPSS переключается между окнами автоматически, в остальных случаях вам потребуется осуществлять переключение вручную.

Когда несколько программ выполняются одновременно, каждой из них соответствует кнопка на панели задач. Ниже показана панель задач с кнопками окна редактора данных и окна вывода.



Чтобы переключиться в окно нужной программы, достаточно подвести указатель к соответствующей кнопке и щелкнуть па ней.

# Создание рабочего каталога

Прежде всего, создайте новую папку, в которой будут храниться создаваемые вами файлы данных и результатов работы с SPSS. В эту папку необходимо скопировать файлы примеров. Предположим, это будет папка с именем C:\SPSSWORK.

Теперь необходимо зарегистрировать эту папку как рабочий каталог (то есть как папку, открываемую по умолчанию) для SPSS.

1. Наведите указатель мыши на значок программы SPSS на рабочем столе и щелкните правой кнопкой мыши.
2. В открывшемся контекстном меню выберите команду Свойства (Properties).
3. Откроется диалоговое окно Свойства: SPSS 11.5 (Properties: SPSS 11.5).
4. Введите в поле Рабочая папка (Working directory) текст: C:\ SPSSWORK.
5. Подтвердите ввод щелчком па кнопке OK.

Рабочий каталог задан. В дальнейшем SPSS будет использовать рабочий каталог как зарегистрированную папку — то есть по умолчанию обращаться к ней в поисках файлов.

# Кнопки

Кнопки используются практически в каждом окне. Они выглядят по-разному и могут находиться как внутри окна, так и в его строке заголовка. Есть несколько стандартных кнопок, которые выполняют одну и ту же функцию независимо от того, окну какой программы они принадлежат (табл. 2.1).

**Таблица 2.1.** Стандартные кнопки программы

| Кнопка | Описание |
|---|---|
| | Кнопка Maximize (Развернуть) находится в верхнем правом углу окна. Если щелкнуть на ней. то окно займет весь экран |
| | Кнопка Minimize (Свернуть) располагается в верхнем правом углу окна слева от кнопки Maximize (Развернуть) и предназначена для свертывания развернутого окна, то есть уменьшения его до размера кнопки на панели задач |
| | Кнопка Restore (Восстановить) уменьшает размеры развернутого окна. Эта кнопка полезна тем, что позволяет видеть на экране одновременно несколько несвернутых окон и управлять ими |
| | Эта кнопка может действовать по-разному в зависимости от использующей ее программы, однако в контексте SPSS она, как правило, предназначена для перемещения элемента (зачастую имени переменной и т. п.) из исходного списка в целевой список |
| | Эта кнопка выполняет действие, обратное предыдущей кнопке, то есть возвращает элемент из целевого списка в исходный список |
| | Эти кнопки предназначены для перемещения в окне с помощью горизонтальной fe. и вертикальной полос прокрутки. Полосы прокрутки появляются в тех случаях, когда размер данных по длине или ширине превышает текущий размер окна. Как правило, для ускорения навигации вместо стрелок используется бегунок прямоугольной формы, расположенный внутри каждой полосы прокрутки |
| | Эта кнопка является обязательным атрибутом раскрывающихся списков и предназначена для их раскрытия |
| | Кнопки счетчика. Обычно счетчик связан с какой-либо числовой величиной, и щелчок на верхней кнопке увеличивает эту величину, а щелчок на нижней кнопке ее уменьшает |

# Окна программы

Существует три типа окон, которые при работе с пакетом SPSS используются чаще других.

- Главное окно программы появляется при запуске SPSS, с него начинается работа с пакетом.
- Диалоговое окно Open File (Открытие файла) позволяет получить доступ к ранее созданным файлам.
- Базовые диалоговые окна хотя и зависят от конкретной процедуры, имеют схожие элементы интерфейса.

Помимо этих трех типов окон особое значение имеет окно вывода. Окно вывода появляется каждый раз после окончания обработки данных. Оно содержит результаты обработки, а также краткие пояснения по их интерпретации. Мы рассмотрим работу с окном вывода и печатью результатов позднее в этой главе, а сейчас обратим внимание на окно редактора данных, или главное окно, которое появляется на экране при запуске SPSS.

## Главное окно

Для того чтобы активизировать главное окно SPSS (рис. 2.1), наведите указатель мыши на кнопку Untitled — SPSS Data Editor (Безымянный — Редактор данных SPSS) панели задач и щелкните па ней (предполагается, что программа SPSS уже запущена).

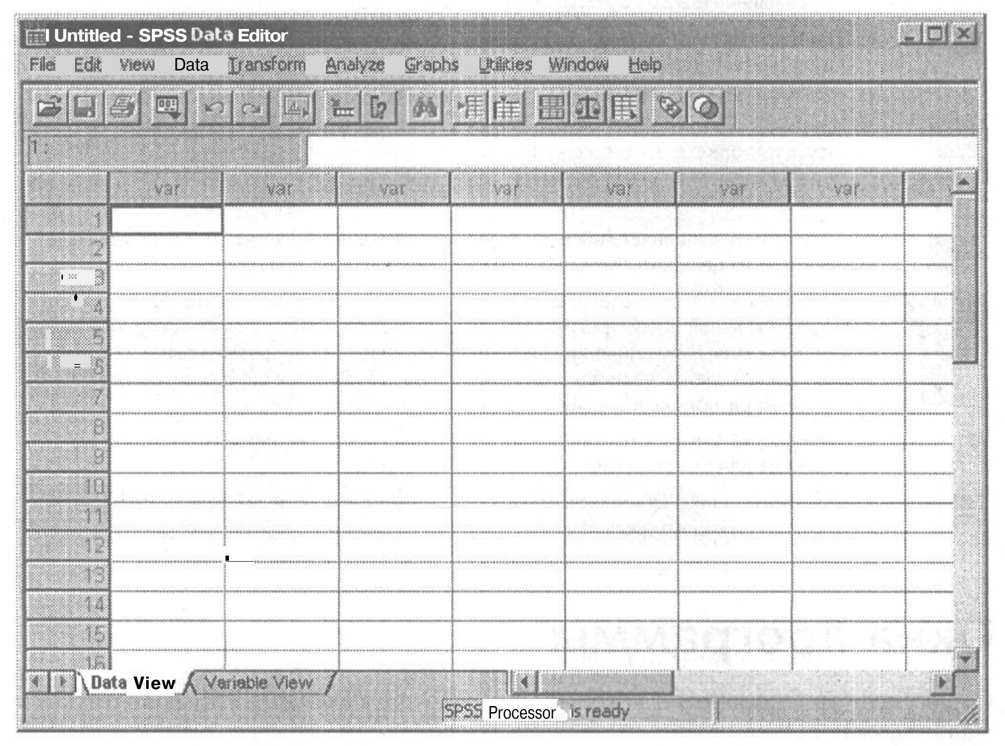



Рис. 2.1. Окно редактора данных SPSS

В верхней части окна расположены строка меню, содержащая команды, и панель инструментов. В момент запуска SPSS редактор данных пуст. Для того чтобы ввести данные, необходимо либо набрать их вручную, заполнив нужные ячейки, либо открыть существующий файл данных. Первый способ подробно описан в главе 3, а в этой главе мы займемся открытием файлов.

Панель инструментов расположена под строкой меню и содержит набор кнопок с различными значками (табл. 2.2). Щелчок на такой кнопке приводит к выполнению некоторой операции. Как правило, панель инструментов позволяет выполнять без помощи меню те операции, которые требуются наиболее часто. Обратите внимание, что в различных приложениях SPSS вид панели инструментов может быть разным. Чтобы убедиться в этом, достаточно сравнить панели инструментов окна данных и окна вывода (см. далее раздел «Окно вывода»).

**Таблица 2.2.** Кнопки панели инструментов редактора данных

| Кнопка | Описание |
|---|---|
| | Открытие файла |
| | Сохранение файла |
| | Печать файла |
| | Повторный вызов последней исполнявшейся команды |
| | Отмена последней операции |
| | Повторение последней отмененной операции |
| | Переход к диаграмме |
| | Переход к объекту (строке) |
| | Отображение информации о текущей переменной |
| | Поиск данных |
| | Вставка нового объекта (строки) в файл данных |
| | Вставка новой переменной в файл данных |
| | Разбиение на подгруппы |
| | Вес объектов (строк) |
| | Выбор объектов (строк) |
| | Переключение между значениями переменных и их метками |
| | Доступ к наборам переменных, определенных пользователем |

Как вы, вероятно, успели заметить, некоторые кнопки выглядят блекло. Это означает, что соответствующая команда в данный момент недоступна пользователю. Так, кнопка Print File (Печать) недоступна потому, что в программе пет данных, которые можно было бы напечатать. Как только вы введете данные, кнопка Print File (Печать) сразу станет доступной. Для того чтобы приобрести навыки работы с панелями инструментов, мы рекомендуем вам самостоятельно «пощелкать» па различных кнопках и понаблюдать за реакцией программы.

Строка меню содержит команды для выполнения почти всех операций, предусмотренных в программе SPSS. По мере изучения этой книги вы познакомитесь с большим количеством команд. Как правило, выполнение команды начинается с появления диалогового окна, в котором пользователю предлагается установить значения параметров. Ниже приводится краткое описание основных меню.

- File (Файл). Содержит команды, предназначенные для открытия, чтения и сохранения файлов, а также команду выхода из программы SPSS.
- Edit (Редактирование). В этом меню находятся команды редактирования, такие как команды копирования, вставки, замены, поиска и т. п.
- View (Просмотр). Содержит набор команд, влияющих на представление информации на экране. Наиболее часто используются команды Value Labels (Метки значений) и Fonts (Шрифты).
- Data (Данные). Здесь находятся команды, предназначенные для управления вводом и представлением данных.
- Transform (Преобразование). Содержит команды, модифицирующие введенные данные, а также создающие новые данные на основе существующих.
- Analyze (Анализ). С этого меню начинаются все процедуры анализа данных.
- Graphs (Графики). Здесь находятся команды, позволяющие создавать различные диаграммы. Иногда с помощью команд этого меню задаются параметры статистических процедур.
- Utilities (Утилиты). Команды этого меню служат для упрощения сложных операций над данными и в основном предназначены для опытных пользователей. Рассмотрение большей части утилит выходит за рамки темы книги.
- Window (Окно). С помощью этого меню можно управлять взаимным расположением и статусом открытых окоп программы SPSS. Фактически, меню Window (Окно) служит средством переключения между окнами, альтернативным панели задач.
- Help (Справка). Как следует из названия меню, оно предназначено для доступа к справочной информации. Весьма полезным при изучении этой книги было бы время от времени обращаться к справочной системе, чтобы расширять получаемые знания.

## Диалоговое окно открытия файла

Диалоговое окно Open File (Открытие файла), представленное на рис. 2.2, позволяет открывать ранее созданные файлы данных. Для того чтобы вызвать это окно, выберите в меню File (Файл) команду Open ▸ Data (Открыть ▸ Данные) либо щелкните мышью па кнопке Open File (Открыть файл) панели инструментов. Содержимое открытого таким образом окна будет соответствовать либо папке, с которой вы работали ранее, либо папке, заданной в качестве рабочего каталога (папка, открываемая программой по умолчанию).

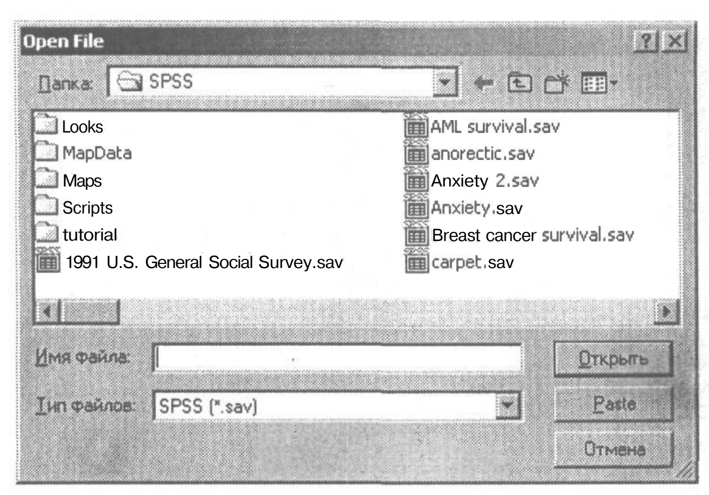



**Рис. 2.2.** Диалоговое окно Open File

Чтобы сделать описание окна более наглядным, воспользуемся файлом данных первого из примеров с именем ex01.sav. Обратите внимание, что все файлы, созданные с помощью редактора данных SPSS, имеют расширение .sav, которое, в зависимости от настройки вашего компьютера, может отображаться или не отображаться в диалоговом окне.

Существует несколько способов открытия файлов. Если нужный вам файл представлен в списке диалогового окна Open File (Открытие файла), то для открытия файла достаточно дважды щелкнуть на нем мышью. Вы также можете ввести имя файла с клавиатуры в поле File Name (Имя файла) и щелкнуть на кнопке Open (Открыть). При этом если файл не находится в текущей папке, необходимо указать полный путь к нему.

Если файл не был создай в редакторе данных SPSS, то перед открытием необходимо указать его тип. Для этого щелкните на стрелке раскрывающегося списка Files of type (Тип файлов ) и выберите тип, соответствующий вашему файлу. К примеру, если файл создавался программой Excel, то нужно выбрать пункт Excel (*.xls). После задания тина файла вы можете воспользоваться одним из двух вышеописанных способов его открытия.

В результате открытия файла часть ячеек окна редактора данных заполнится данными, которые можно обрабатывать с помощью команд меню.

## Диалоговое окно процедуры обработки

Каждая процедура обработки имеет собственное диалоговое окно. Несмотря на это практически все диалоговые окна построены по одному и тому же принципу. Мы продемонстрируем это на примере диалогового окна процедуры Frequencies (Частоты), представленного на рис. 2.3 (см. также главу 6).

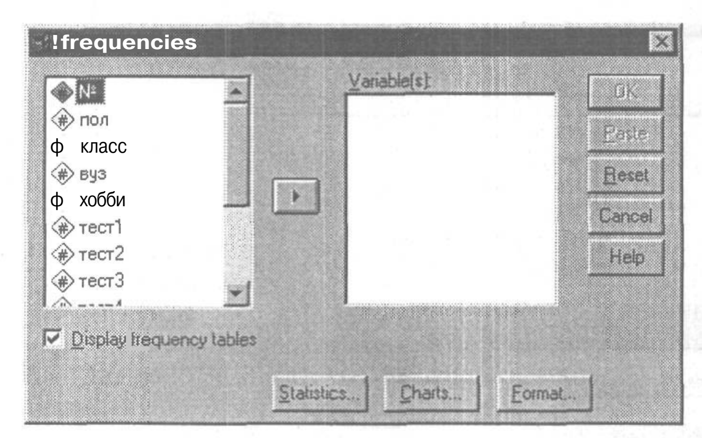



**Рис. 2.3.** Диалоговое окно Frequencies

В левой части окна расположен список, содержащий все доступные переменные файла, справа могут находиться от одного до трех изначально пустых списков. Несмотря на то что заголовки этих списков различны, все они предназначены для указания переменных, участвующих в процедуре обработки. Между списками находится кнопка со стрелкой, позволяющая перемещать переменные из одного списка в другой. Стрелка на кнопке соответствует направлению перемещения нeременных. Если выделить пункт в левом списке, стрелка будет направлена вправо, если в правом списке — влево.

В диалоговых окнах всех процедур обработки имеется пять кнопок.

- Кнопка OK инициирует процедуру обработки выбранных переменных.
- Кнопка Paste (Вставка) открывает окно, в котором отображается синтаксическая конструкция, соответствующая вызову команды с параметрами, указанными в списках диалогового окна (эту конструкцию называют командным файлом, поскольку она записывается в специальный файл, который при желании можно сохранить и использовать в дальнейшем). Непосредственная работа с командным файлом предоставляет более широкие возможности, чем настройка параметров команды с помощью элементов управления диалогового окна. Как правило, подобные возможности доступны только опытным пользователям в специфических случаях и требуют определенных программистских навыков. Если вы не обладаете такими навыками, то либо внимательно изучите хотя бы несколько сотен страниц синтаксического справочника, либо на время забудьте о существовании этой кнопки.
- Кнопка Reset (Сброс) возвращает параметры к значениям по умолчанию. Каждый раз при проведении обработки набор исходных данных сохраняется, что позволяет использовать его многократно с разными значениями параметров. Это удобно, если вы проводите серию статистических опытов с одними и теми же переменными. Если же вам необходимо изменить начальные условия, то, щелкнув на кнопке Reset (Сброс), вы сможете очистить список переменных и установить значения параметров, принятые по умолчанию.

- Кнопка Cancel (Отмена) позволяет немедленно закрыть окно без выполнения процедуры обработки.
- Кнопка Help (Помощь) открывает доступ к *контекстной* помощи. Это означает, что если вы щелкнете на этой кнопке в окне Frequencies (Частоты), то получите описание именно процедуры Frequencies (Частоты), а не общую справку по всем процедурам обработки.

Наконец, кнопки Statistics (Статистики), Charts (Диаграммы) и Format (Формат), находящиеся в нижней части диалогового окна, позволяют задавать различные параметры, управляющие представлением результатов процедуры и ее выполнением. Почти все диалоговые окна имеют подобные кнопки. Число кнопок и параметры, которые они позволяют задавать, меняются в зависимости от конкретной процедуры.

## SPSS без мыши

До сих пор существуют пользователи, которые по тем или иным причинам отказываются от использования мыши. Как правило, программное обеспечение разрабатывается таким образом, что все действия (или их основная часть) были выполнимы при помощи клавиатуры. Пакет SPSS не является исключением, и мы, не вдаваясь в подробности, рассмотрим, каким образом управлять его работой без мыши.

Вместо щелчка мыши на выделенном элементе достаточно просто нажать клавишу Enter. Для того чтобы выделить элемент с помощью клавиатуры, можно использовать клавиши табуляции, пробела, а также клавиши со стрелками. Выбор пункта меню осуществляется двумя способами: перемещением выделения от пункта к пункту либо с помощью клавиш со стрелками, либо с помощью комбинаций клавиш. Как вы можете видеть, в большинстве названий команд одна из букв подчеркнута. Это означает, что если нажать клавишу Alt в сочетании с этой буквой, то будет выполнена соответствующая команда. Необходимо помнить, что комбинация клавиш действует только для активного в данный момент меню. Описанный способ запуска команд имеет своих сторонников, и их число сравнимо с теми, кто для работы с командами предпочитает использовать мышь. То же самое можно сказать и о выделении текста с помощью клавиатуры.

Чтобы выделить блок ячеек, следует выделить угловую ячейку намеченного блока, а затем, удерживая клавишу Shift, с помощью клавиш со стрелками расширить выделение. Существуют и другие варианты использования клавиатуры в программе SPSS, однако они гораздо менее важны, и вам будет достаточно полученных знаний для того, чтобы успешно обходиться без мыши при изучении этой книги.

Помимо кнопок при работе с диалоговыми окнами вы можете столкнуться еще с двумя типами элементов управления. Обратимся к рис. 2.4, на котором представлено диалоговое окно, состоящее из двух областей. Первая область представляет собой группу *флажков,* каждый из которых может быть установлен или сброшен. Если вы устанавливаете флажок, то рядом с его меткой появляется «галочка».

Если галочка отсутствует, то флажок полагается сброшенным. Для изменения состояния флажка необходимо щелкнуть на месте расположения галочки. Как правило (хотя и не всегда), флажки независимы, то есть установка или сброс одного из них никак не влияет на состояние остальных.

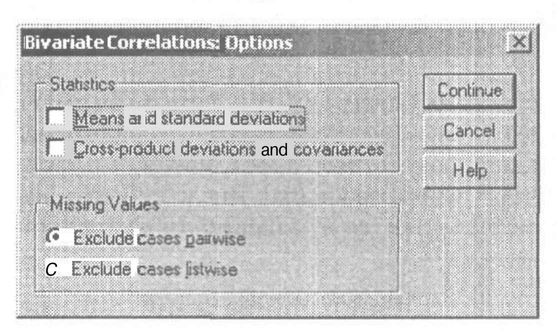



**Рис. 2.4.** Диалоговое окно с флажками и переключателями

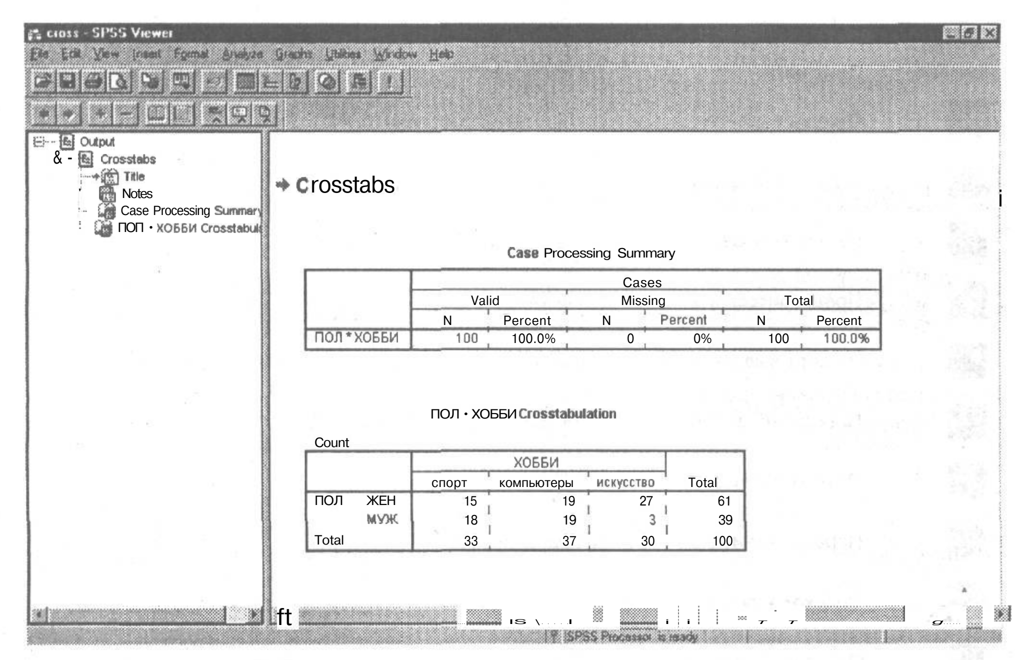
Вторая область представляет собой группу из двух переключателей. Переключатели связаны друг с другом таким образом, что если один из них установлен, то остальные переключатели в группе полагаются сброшенными. Программные переключатели по своему смыслу аналогичны физическим переключателям: если, к примеру, на магнитофоне включен режим проигрывания, то режимы перемотки и записи оказываются выключенными. Установленный переключатель обозначается точкой в соответствующем кружке.

# Окно вывода

Под термином «вывод» в этой книге, как правило, понимается совокупность результатов любой процедуры обработки. Задача представления вывода в максимально простой и понятной форме по-разному решалась в разных версиях программы SPSS. В версиях 6.x для вывода использовался шрифт Courier кегля 10, а длина каждой строки умещалась в пределы страницы. В данной версии используется табличный формат. Пример окна вывода представлен на рис. 2.5. Этот формат имеет заметные преимущества, однако значительный объем выводимой информации иногда неудобен при чтении.

Каждая глава, в которой упоминается вывод данных, обязательно снабжена указаниями о содержании окна вывода. Поскольку это окно содержит главный результат обработки данных, уделим особое внимание возможности его редактирования. Конечно, можно работать с выходными данными, распечатанными на бумаге, и не менять формат вывода перед печатью, однако следует иметь в виду следующие моменты.

- При выводе больших объемов данных требуется большое количество листов бумаги, причем нередко пространство печати используется неэффективно.
- Часть выводимой информации, как правило, оказывается ненужной для вас.




| | Cases | | | | | |
|---|---|---|---|---|---|---|
| | Valid | | Missing | | Total | |
| | N | Percent | N | Percent | N | Percent |
| ПОЛ * ХОББИ | 100 | 100.0% | 0 | 0% | 100 | 100.0% |

| | | ХОББИ | | | Total |
|---|---|---|---|---|---|
| | | спорт | компьютеры | искусство | |
| ПОЛ | ЖЕН | 15 | 19 | 27 | 61 |
| | МУЖ | 18 | 19 | 3 | 39 |
| Total | | 33 | 37 | 30 | 100 |

**Рис. 2.5.** Окно вывода SPSS

- Большие таблицы, создаваемые программой SPSS, иногда неудобны для восприятия, поэтому желательна их реорганизация.
- Наличие комментариев в выводимых данных впоследствии облегчит их понимание для вас, а также для людей, использующих результаты обработки.

Рассмотрим подробно возможности работы с окном вывода SPSS и его панелью инструментов. Это позволит вам в дальнейшем удалять ненужные фрагменты полученных данных, добавлять в них свои комментарии, реорганизовывать порядок вывода и сохранять данные в файле.

Некоторые кнопки па панели инструментов окна вывода совпадают с кнопками окна редактора данных. Эти кнопки действительно выполняют аналогичные действия; разница состоит лишь в том, что в редакторе данных обрабатываются исходные данные, а в окне вывода — результаты (табл. 2.3).

Строго говоря, окно вывода на самом деле не является окном. Как видно на рис. 2.5, результаты анализа отображены в окне Output — SPSS Viewer (Вывод — Просмотр SPSS) — специальной программы просмотра результатов анализа. Окно Output — SPSS Viewer (Вывод — Просмотр SPSS) разделено на две области. Правая область представляет собой то, что мы назвали окном вывода, то есть окно результатов анализа данных. В левой области отображена иерархическая структура объектов окна вывода. В сущности, весь вывод программы — заголовки, таблицы, диаграммы и т. п., на самом деле являются не чем иным, как разного рода объектами, имеющими иерархическую структуру. Таким образом, вместе с результатами анализа мы видим их внутреннее программное представление.

**Таблица 2.3.** Кнопки панели инструментов окна вывода

| Кнопка | Описание |
|---|---|
| | Открытие файла |
| | Сохранение файла |
| | Печать вывода |
| | Предварительный просмотр печати |
| | Экспорт вывода в текстовый файл (Word/RTF. Text и т. п.) или веб-страницу |
| | Повторный вызов последней исполнявшейся команды |
| | Отмена последней операции |
| | Переключение в окно редактора данных SPSS |
| | Переход к объекту (строке) данных |
| | Отображение информации о текущей переменной |
| | Доступ к наборам переменных, определенных пользователем |
| | Переход в конец вывода последней выполненной процедуры |
| | Определение окна, в которое будет направлен вывод |
| | Повышение уровня выделенного заголовка |
| | Понижение уровня выделенного заголовка |
| | Развертывание содержимого заголовка |
| | Свертывание содержимого заголовка |
| | Отображение текущего скрытого объекта |
| | Скрытие текущего отображенного объекта |
| | Вставка заголовка над надписями, таблицами и диаграммами |
| | Вставка названия в выделенный объект |
| | Вставка текста после выделенного объекта |

Если вы внимательно посмотрите на содержимое окна Output — SPSS Viewer (Вывод — Просмотр SPSS), то заметите, что объект Notes, отображенный в иерархической структуре, отсутствует в окне вывода. Это объясняется тем, что по умолчанию он является *скрытым*. Слева от объектов расположены значки, принимающие вид открытой или закрытой книги в зависимости от того, скрыты или отображены эти объекты в данный момент. Если дважды щелкнуть на таком значке, состояние объекта изменится на противоположное.

Иерархическая структура делает удобной навигацию внутри окна вывода. Так, если вам нужно просмотреть содержимое какой-либо таблицы, достаточно щелкнуть па объекте, чтобы соответствующее изображение появилось на экране. Как нетрудно догадаться, операция удаления фрагментов вывода сводится к удалению соответствующих объектов. Для удаления объекта щелкните па его имени в левой части окна, а затем в меню Edit (Редактирование) выберите команду Delete (Удаление). Если вы хотите изменить порядок следования объектов, то выделите один или несколько объектов, которые намереваетесь переместить, а затем в меню Edit (Редактирование) выберите команду Cut (Вырезание). После этого выделите объект, за которым вы хотите осуществить вставку, и выберите команду Edit ▸ Paste After (Редактирование ▸ Вставить после).

Уже после нескольких операций обработки одних и тех же данных навигация внутри окна вывода становится затруднительной даже при иерархической структуре объектов. Для решения проблемы используется возможность сворачивания объектов. Как можно видеть на рис. 2.6, слева, около названий объектов имеется значок - (минус). Если вы не хотите, чтобы объекты под каким-либо из заголовков присутствовали на экране, щелкните на этом значке. После этого названия соответствующих объектов исчезнут, на схеме останется лишь заголовок, а вместо значка - (минус) появится значок + (плюс), как показано на рис. 2.6 справа. Для того чтобы снова развернуть список объектов, достаточно щелкнуть на этом значке, и содержимое схемы будет восстановлено.

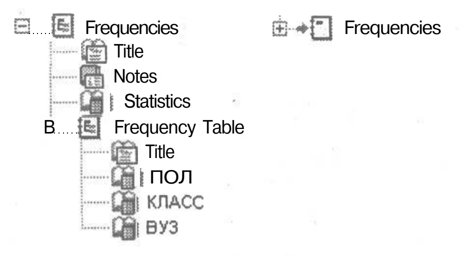



**Рис. 2.6.** Развернутая (слева) и свернутая иерархии

Еще одной удобной возможностью при работе с пакетом SPSS является возможность вставки текста в выводимые данные. Это делается при помощи команды Insert ▸ New Text (Вставка ▸ Текст). Фактически данная команда позволяет вставить в окно вывода текстовый объект SPSS Text. Наличие текстовых комментариев в выводимых данных программы упрощает понимание полученных результатов,

особенно людьми, не принимавшими участия в подготовке и обработке данных. После того как нужные комментарии введены, щелкните па любом другом объекте, чтобы снять выделение с объекта SPSS Text.

Теперь, когда вы получили элементарные знания о том, что представляет собой окно вывода программы SPSS и каким образом работать с ним на объектном уровне, пришло время заглянуть внутрь самих объектов и научиться изменять их. Нас будет интересовать структура таблиц и диаграмм, поскольку от того, насколько удобно представлены эти элементы вывода, зависит удобочитаемость выводимых данных в целом. В этой книге мы не будем рассматривать все процедуры, изменяющие представление таблиц и диаграмм, однако знаний, которые вы получите, вполне достаточно для комфортной работы с результатами исследований. Редактирование диаграмм описывается в главе 5.

В первую очередь необходимо уяснить принцип построения таблиц. Обычно SPSS создает таблицы автоматически, самостоятельно решая, какие показатели разместить в строках, а какие — в столбцах. Конечно, не всегда удобно полагаться на выбор программы, и для этих случаев SPSS предлагает средства редактирования таблиц.

Чтобы попрактиковаться в редактировании таблиц, возьмем результат обработки из главы 8 — таблицу сопряженности пол×хобби×класс. Все описываемые операции вы можете проделать сами, если воспользуетесь файлом примера cross.spo. В таблице сопряженности (из файла cross.spo), представленной на рис. 2.7, SPSS по умолчанию размещает данные о хобби учащихся в столбцах, а данные по полу и классу — в строках. Несмотря па то что в подобном представлении нет ничего неправильного, может возникнуть ситуация, когда для наглядности потребовалось бы расположить данные в таблице по-другому. При этом сами данные необходимо оставить неизменными, поскольку они являются результатом анализа.

ПОЛ * ХОББИ * КЛАСС Crosstabulation

Count

| | | | ХОББИ | | | |
|---|---|---|---|---|---|---|
| КЛАСС | | | спорт | компьютеры | искусство | Total |
| 1.00 | ПОЛ | ЖЕН | 6 | 5 | 3 | 14 |
| | | МУЖ | 11 | 7 | 1 | 19 |
| | Total | | 17 | 12 | 4 | 33 |
| 2.00 | ПОЛ | ЖЕН | 6 | 6 | 9 | 21 |
| | | МУЖ | 5 | 7 | 2 | 14 |
| | Total | | 11 | 13 | 11 | 35 |
| 3.00 | ПОЛ | ЖЕН | 3 | 8 | 15 | 26 |
| | | МУЖ | 2 | 4 | 0 | 6 |
| | Total | | 5 | 12 | 15 | 32 |

**Рис. 2.7.** Исходная таблица сопряженности

В столбцах рассматриваемой таблицы перечислены хобби учащихся (спорт, компьютеры, искусство), а каждая строка содержит два уровня заголовков: номер класса (1, 2, 3) и пол ученика (ЖЕН, МУЖ). При этом для каждого заголовка первого уровня перебраны всевозможные варианты заголовков второго уровня. Данная таблица является однослойной, а число предметов в каждой категории (Count) отображено в верхней левой части таблицы. Большинство используемых на практике таблиц также являются однослойными, поэтому мы сразу перейдем к организации столбцов и строк.

Для того чтобы производить какие-либо действия с таблицей, сначала необходимо дважды щелкнуть на ней в окне вывода. В ответ на экране исчезнут напели инструментов, а в строке меню появится новое меню Pivot (Вращение). Кроме того, изменится набор доступных команд меню. Далее будут рассмотрены команды Rotate inner column labels (Вращать внутренние метки столбцов), Rotate outer row labels (Вращать наружные метки строк), Transform rows and columns (Переместить строки и столбцы) и Pivoting trays (Места осей) меню Pivot (Вращение) и Format (Формат).

Вероятно, вы уже обратили внимание на то, что ширина столбцов рассматриваемой таблицы гораздо больше, чем необходимо для отображения содержащихся в ней чисел. Это объясняется тем, что дополнительное пространство занято именами (метками) столбцов. Если изменить направление текста меток столбцов, то есть применить к таблице команду Format ▸ Rotate inner column labels (Формат ▸ Вращать внутренние метки столбцов), то формат таблицы станет гораздо более компактным (рис. 2.8).

ПОЛ * ХОББИ * КЛАСС **Crosstabulation**

Count

| | | | ХОББИ | | | |
|---|---|---|---|---|---|---|
| КЛАСС | | | спорт | компьютеры | искусство | Total |
| 1.00 | ПОЛ | ЖЕН | 6 | 5 | 3 | 14 |
| | | МУЖ | 11 | 7 | 1 | 19 |
| | Total | | 17 | 12 | 4 | 33 |
| 2.00 | ПОЛ | ЖЕН | 6 | 6 | 9 | 21 |
| | | МУЖ | 5 | 7 | 2 | 14 |
| | Total | | 11 | 13 | 11 | 35 |
| 3.00 | ПОЛ | ЖЕН | 3 | B | 15 | 26 |
| | | МУЖ | 2 | 4 | 0 | 6 |
| | Total | | 5 | 12 | 15 | 32 |

**Рис. 2.8.** Таблица сопряженности после поворота меток столбцов

Для того чтобы восстановить предыдущий формат, достаточно выбрать указанную команду еще раз. Аналогичным образом можно изменять формат строк, только в меню Format (Формат) выбирать команду Rotate outer row labels (Вращать наружные метки строк).

Как уже говорилось, при создании таблицы содержимое ее строк и столбцов назначается программой автоматически. Если вам необходимо изменить расположение категорий в таблице, то следует воспользоваться командой Transform rows and columns (Переместить строки и столбцы) или Pivoting trays (Места осей). Первая меняет местами строки и столбцы; результат ее применения к рассматриваемой таблице показан па рис. 2.9.

ПОЛ * ХОББИ * КЛАСС **Crosstabulation**

Count

| | | КЛАСС | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1.00 | | | 2.00 | | | 3.00 | | |
| | | ПОЛ | | Total | ПОЛ | | Total | ПОЛ | | |
| | | ЖЕ | МУЖ | | ЖЕ | МУЖ | | ЖЕ | МУЖ | |
| ХОББИ | спорт | 6 | 11 | 17 | 6 | 5 | 11 | 3 | 2 | 5 |
| | компьютеры | 5 | 7 | 12 | 6 | 7 | 13 | 8 | 4 | 12 |
| | искусство | 3 | 1 | 4 | 9 | 2 | 11 | 15 | 0 | 15 |
| Total | | 14 | 19 | 33 | 21 | 14 | 35 | 26 | 6 | 32 |

**Рис. 2.9.** Таблица сопряженности после перестановки строк и столбцов

Вторая команда позволяет осуществить более гибкую реорганизацию таблицы. При ее выборе на экране появляется новое окно Pivoting Trays (Места осей), представленное на рис. 2.10, в котором вы можете задать содержимое будущих строк и столбцов. Команда Pivoting trays (Места осей) также предназначена для реорганизации слоев, но поскольку необходимость в такой операции возникает крайне редко, мы не будем ее рассматривать.

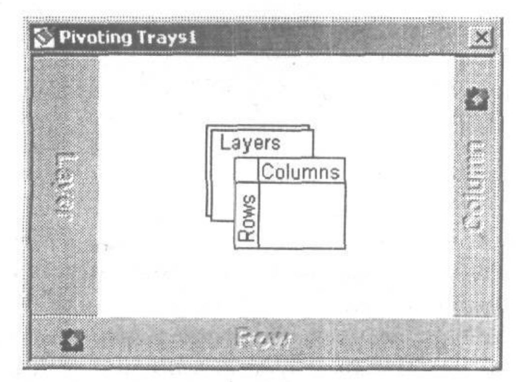



**Рис. 2.10.** Окно Pivoting Trays

Реорганизация таблицы производится с помощью *осей*. Под *осью строк* будем понимать столбец, содержащий метки строк, а под *осью столбцов* — строку, содержащую метки столбцов. В нашем примере на оси строк находится информация о номере группы и поле учащихся, а на оси столбцов — об их хобби.

В окне Pivoting Trays (Места осей) имеются панели Row (Строка) и Column (Столбец), соответствующие оси строк и оси столбцов. Вы можете менять положение меток, перетаскивая квадратные значки из разноцветных стрелок в пределах одной оси или с одной оси на другую. Если вы хотите, к примеру, поменять местами класс и пол учащихся, достаточно перетащить значок пола, расположив его левее значка класса. В этом случае пол окажется заголовком первого уровня, и каждая метка пола будет содержать три метки класса.

Вы также можете реорганизовывать переменные, перемещая соответствующие квадратные значки между осями строк и столбцов. Так, если вам необходимо сделать номер класса переменной первого уровня, для каждого значения которой отображались бы всевозможные варианты хобби, то следует перетащить переменную номера класса выше переменной хобби.

Как вы уже успели заметить, даже для такой простой таблицы, как эта, существует большое количество вариантов организации. Самый простой способ научиться реорганизовывать таблицы — поэкспериментировать с ними, воспользовавшись примером cross.spo или своими результатами. Если вы запутаетесь в своих действиях, вы всегда можете отменить результаты и начать заново.

Иногда выводимые данные занимают слишком много места и представлены в окне вывода лишь частично. В этом случае в том месте окна вывода, в котором скрыты дополнительные данные, присутствует значок в виде красного треугольника. Чтобы увидеть итоговые данные полностью, дважды щелкните на представленном в окне вывода фрагменте. На рис. 2.11 показана часть окна вывода со скрытыми данными (о наличии скрытых данных сигнализирует треугольный значок слева).

| | | | | |
|---|---|---|---|---|
| .00 | | | | |
| ХОББИ | компьюте | 2.5 | 2.5 | .00 |
| ВУЗ | ЭКОН | | | |
| ХОББИ | спорт | .5 | .5 | .00 |

**Рис. 2.11.** Часть окна вывода со скрытыми данными

После того как формат выводимых данных программы будет изменен в соответствии с вашими предпочтениями, у вас, вероятно, возникнет необходимость сохранить результат в файле для дальнейшего использования. В отличие от файла данных, файл вывода имеет расширение .spo. Для сохранения файла вывода в меню File (Файл) выберите команду Save (Сохранение), в появившемся диалоговом окне задайте имя файла и щелкните на кнопке Save (Сохранить).

Начиная с версии 11.5, программа SPSS предоставляет возможность сохранения файлов вывода и в более привычном для многих формате текстового редактора Word. Для этого в меню File (Файл) выберите команду Export (Экспорт). На экране появится диалоговое окно Export Output (Экспорт файла вывода), показанное на рис. 2.12. В этом окне необходимо в списке File type (Тип файла) выбрать пункт Word/RTF file (*.doc), в поле Export File (Файл экспорта) ввести имя нового файла и задать место его расположения, щелкнув на кнопке Browse (Просмотр). Дополнительно с помощью списка Export (Экспорт) можно указать, нужно ли экспортировать графики: Output document (Весь документ), Output document (No Charts) (Весь документ без графиков), Charts only (Только графики), а также с помощью группы переключателей Export What (Экспортируемые объекты) выбрать объекты для экспорта: All Objects (Все объекты), All Visible Objects (Все видимые объекты), Selected Objects (Выделенные объекты). После окончания настройки в этом диалоговом окне щелкните на кнопке ОК. Будет создан файл с расширением .doc, содержащий те объекты файла вывода, которые вы выбрали. Этот файл можно будет открывать и редактировать в программе Word.

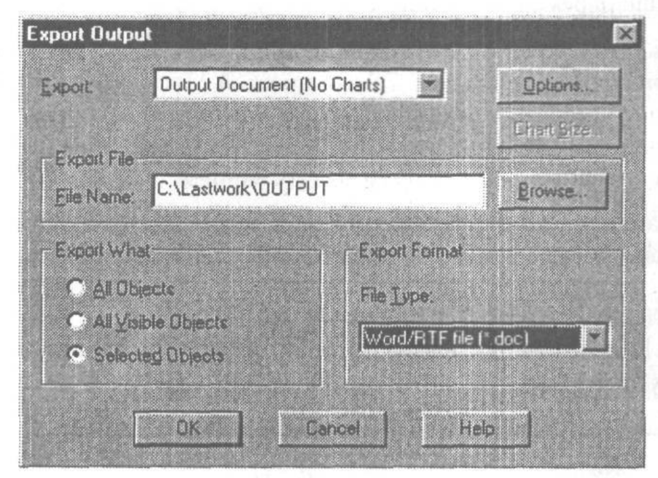



Рис. 2.12. Диалоговое окно Export Output

Другой способ сохранения содержимого окна вывода результатов — копирование в буфер обмена элементов окна вывода (таблиц, графиков) в качестве объектов с помощью команды Copy objects (Копировать объекты) и последующая их вставка в открытый документ Word.

# Печать результатов

Перед тем как начать печать, следует выбрать в меню File (Файл) команду Page Setup (Установка страницы). На экране появится диалоговое окно, в котором можно задать ориентацию страницы (книжную или альбомную), ширину полей, а также размер бумаги и тип устройства ее подачи.

Когда параметры страницы заданы, необходимо выделить фрагменты выводимых данных, однако если вы намерены отправить на печать все данные, то никакого выделения делать не требуется. Мы настоятельно рекомендуем вам перед началом печати обдумать, стоит ли печатать те или иные фрагменты, поскольку искать полезную информацию среди большого объема ненужных таблиц, как правило, достаточно сложно. Чтобы выделить фрагмент выводимых данных, выполните любое из трех перечисленных ниже действий.

- Щелкните на заголовке или объекте иерархической структуры, расположенной в левой части окна вывода. При щелчке на заголовке выбранными оказываются все объекты, расположенные под этим заголовком.
- Тем же способом, описанным в предыдущем пункте, выделите один объект, а затем, удерживая клавишу Shift, выделите другой объект. В результате выбранными окажутся все объекты, расположенные между двумя указанными объектами.
- Способом, описанным в первом пункте, выделите один объект, а затем, удерживая клавишу Ctrl, поочередно выделите все объекты, предназначенные для печати. Этот способ отличается от предыдущего тем, что позволяет выбирать несмежные объекты.

После выделения требуемых объектов в меню File (Файл) выберите команду Print (Печать). На экране появится диалоговое окно, представленное на рис. 2.13. С помощью этого окна можно выбрать принтер (в случае если на компьютере установлено более одного принтера), а также диапазон печати, определяющий, следует ли печатать все данные или только выделенные фрагменты. Когда все параметры установлены нужным образом, щелкните на кнопке OK для завершения процесса.

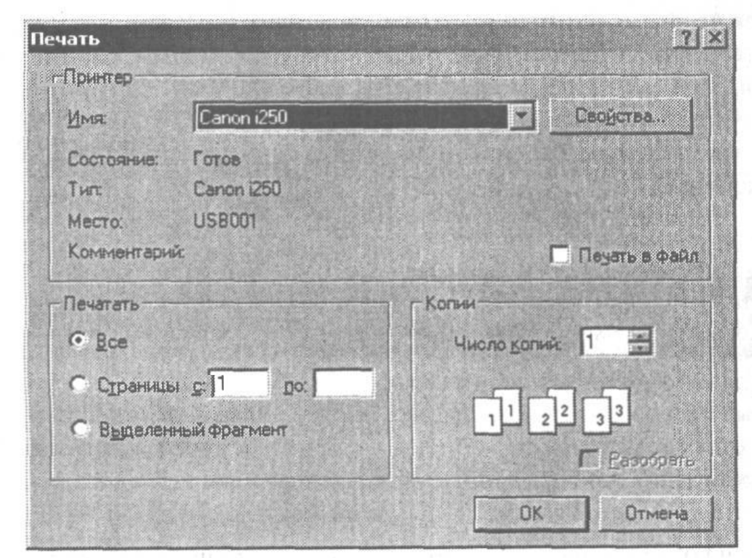



**Рис. 2.13.** Диалоговое окно печати

# 3 Создание и редактирование файлов данных



Эта глава поможет вам сделать первый шаг к статистическому анализу — ввести данные в программу. Как вы уже знаете из предыдущей главы, в SPSS для управления данными используется специальная программа — редактор данных. Окно редактора данных представляет собой электронную таблицу, в которой столбцы отражают переменные, а строки — объекты (случаи). На пересечении строки и столбца находится ячейка, в которой сохраняется значение переменной (столбца) для данного объекта (строки). Для работы с таблицами, не умещающимися в пределах окна, предусмотрены вертикальные и горизонтальные полосы прокрутки, а также автоматические функции перехода к заданной переменной или объекту.

Данная глава является первой в книге, содержащей пошаговые процедуры. В них приведена последовательность действий, необходимых для ввода данных в компьютер. В качестве примера мы будем использовать файл ex01.sav, упоминавшийся в главе 2 и содержащий разнообразную информацию о 100 школьниках, обучающихся в трех классах. Вы можете загрузить этот файл с веб-сайта http://www.piter.com/download.

## Структура файла данных

Перед тем как начать ввод данных, необходимо определить структуру будущего файла. Для этого вы должны ответить себе на вопрос, как будут использоваться в анализах те или иные переменные. К сожалению, многие пользователи начинают задумываться об этом гораздо позже, чем следует, и в этом кроется причина многих неудач при проведении анализа данных: чем сложнее план исследования, тем больше шансов, что из-за необдуманных действий исследователя он завершится неудачей.

Итак, первое, что следует сделать, — определить последовательность действий при обработке данных. Кроме того, необходимо четко представлять себе струк-

туру и взаимосвязи переменных в вашем плане исследования. Ниже перечислены наиболее характерные для файлов данных ошибки и недостатки.

- Отсутствуют ключевые переменные (пол, возраст и т. п.), являющиеся основой для анализа.
- Переменная плохо отражает содержание соответствующей реальной величины (например, на сложный вопрос имеются только два варианта ответа: «да» и «нет»).
- При большом количестве независимых переменных отсутствуют зависимые переменные, отражающие цель исследования (или наоборот).
- Недостаточно независимых переменных, влияющих на заданную зависимую переменную.

Эти примеры демонстрируют, что залогом успеха как исследования в целом, так и создания файла данных в частности является тщательно продуманный выбор структуры данных.

Следует отметить, что структура файла данных должна соответствовать плану исследования. С другой стороны, план исследования должен быть составлен так, чтобы его исходные данные можно было бы обработать в соответствии с задачами и гипотезами исследования. Самый оптимальный и простой путь обеспечения этих соответствий — определение структуры данных на этапе планирования исследования, еще до сбора данных. Это позволит избежать большинства типичных ошибок, относящихся как к планированию исследования, так и к организации данных.

На этапе планирования исследования структура данных может быть задана в виде предварительного списка переменных с указанием их типов и диапазонов возможных значений, например так, как в табл. 3.1.

**Таблица 3.1.** Предварительный список переменных

| № | Название | Тип | Диапазон возможных значений |
|---|---|---|---|
| 1 | Идентификационный номер | Номинативная | 1-100 |
| 2 | Пол | Номинативная | 1-2 |
| 3 | Класс | Номинативная | 1-3 |
| 4 | Предполагаемый для поступления вуз | Номинативная | 1 — гуманитарный; 2 — экономический; 3 — технический; 4 — естественнонаучный |
| 5 | Внешкольные увлечения | Номинативная | 1 — спорт; 2 — компьютер; 3 — искусство |
| 6–10 | Показатели тестов 1–5 | Количественные | 1–20 |
| 11 | Средний балл отметок за 10-й класс | Количественная | 3-5 |
| 12 | Средний балл отметок за 11-й класс | Количественная | 3-5 |

Каждая переменная — это имеющее значение для исследователя основание, позволяющее отличать объекты друг от друга. На предварительном этапе следует выделять два типа нерeменных: количественные и категориальные (номинативные). Количественная переменная позволяет различать объекты по уровню выраженности некоторого свойства, например: средний балл отметки, тестовый показатель и пр. Идентификация количественных переменных на предварительном этапе не составляет труда: обычно они соответствуют тому, что исследователь намеревается измерить. Второй тип — категориальные (номинативные) переменные. Обычно они используются как основания для деления объектов (испытуемых) на группы или категории: пол, класс, возрастная категория, уровень дохода и пр. Типичная ошибка начинающего исследователя — игнорирование возможных оснований для деления объектов на группы в качестве самостоятельных номинативных переменных в структуре данных.

Важным свойством номинативных переменных является возможность их представления в виде набора целых чисел. Например, трем видам внешкольных увлечений (хобби) учащихся (спорт, компьютер, искусство) можно сопоставить числа 1, 2 и 3 соответственно. Числовое представление данных в компьютерных программах всегда предпочтительнее символьного, поскольку обработка чисел происходит быстрее, проще и с меньшей вероятностью ошибок. Кроме того, числовое представление легко модифицировать: вы можете переназначить числа, соответствующие созданным элементам, а также (что часто требуется на практике) без проблем включить в анализ новые элементы. Например, если в группе окажется учащийся, увлечение которого не соответствует перечисленным, будет полезно включить в переменную хобби категорию с названием другие и присвоить ей число 4. Эта операция рассмотрена подробнее в главе 4.

Порядок создания переменных также важен при вводе данных. Здесь следует придерживаться простого правила: наиболее важные и часто используемые переменные должны помещаться в начало файла, для остальных данных вопрос порядка следования не столь важен, однако рекомендуется объединять их в группы по «физическому смыслу». Чаще всего в начало файла следует поместить категориальные переменные, которые далее предполагается использовать для деления объектов (испытуемых) на группы, например пол, семейное положение и пр. Далее можно перечислять остальные сведения, а логическое объединение переменных производить в зависимости от того, какие аспекты они отражают.

Далее мы приведем инструкции, с помощью которых вы сможете приступить к освоению программы SPSS.

**Шаг 1** Первое, что необходимо сделать, — это запустить программу SPSS и в открывшемся диалоговом окне щелкнуть на кнопке Cancel (Отмена). Вы получите доступ к окну редактора данных, представленному на рис. 3.1.

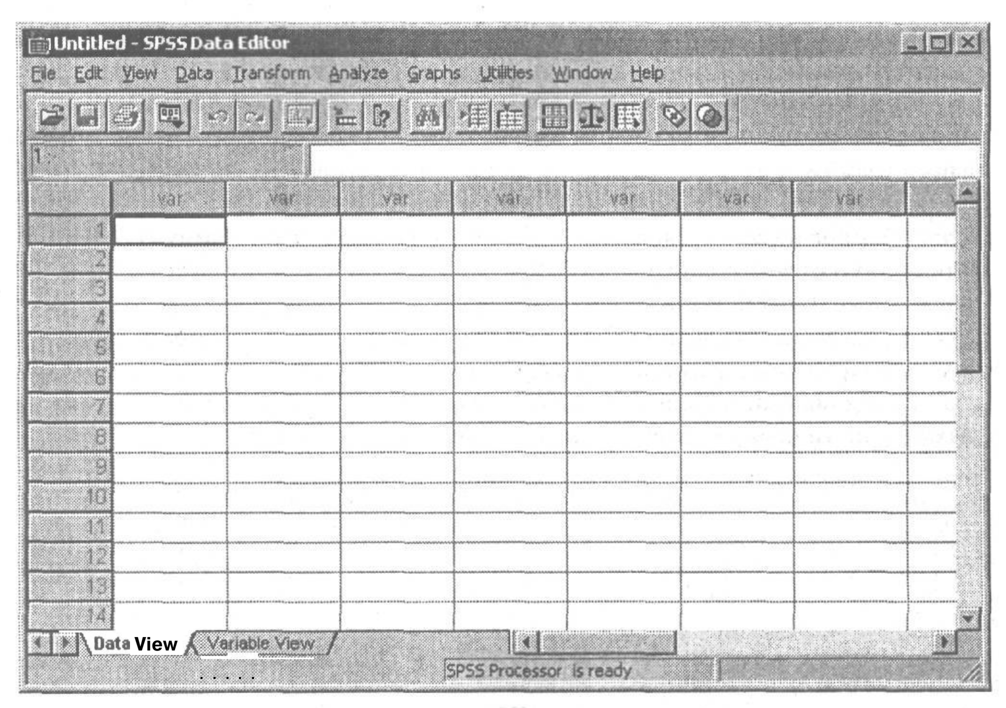



Рис. 3.1. Исходный вид окна редактора данных SPSS

Шаг 2 Перейдите на вкладку Variable View (Просмотр переменных), щелкнув на ее ярлычке мышью (рис. 3.2).

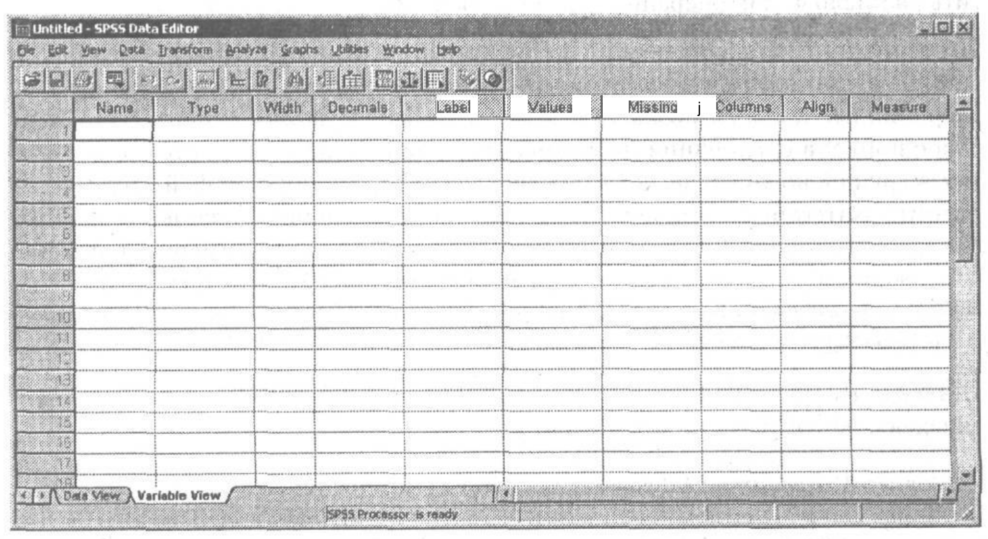



Рис. 3.2. Вкладка просмотра переменных окна редактора данных

Вкладка Data View (Просмотр данных), которая отображается сразу после запуска редактора данных, предназначена для ввода значений в создаваемый файл данных. Вкладка Variable View (Просмотр переменных) позволяет задать структуру файла данных, то есть определить имена, метки и структуры переменных. Заголовки столбцов представляют собой параметры каждой из переменных: Name (Имя), Type (Тип), Width (Ширина), Decimals (Дробная часть), Label (Метка), Values (Значения), Missing (Пропуски), Columns (Столбцы), Align (Выравнивание) и Measure (Измерение). Мы будем описывать процесс создания файла данных в том порядке, в котором только что перечислили параметры переменных.

В следующих пошаговых инструкциях мы возьмем за основу данные из файла ex01.sav. Часть содержимого этого файла приведена в конце этой главы. В процессе создания файла можно выделить три основных действия: задание имени переменной, определение ее параметров и ввод данных. Необязательно последовательно выполнять каждый из трех этапов для каждой переменной, такой порядок выбран лишь в качестве примера. На практике вам может оказаться удобнее создавать файл данных «порциями»: сначала вы полностью зададите часть переменных, обработаете их, затем введете новую группу переменных и т. д. Мы же сейчас обратимся к рассмотрению параметров переменных.

## Имя переменной

Параметр Name (Имя) определяет имя переменной. Чтобы задать имя первой переменной, просто введите его с клавиатуры в текущую ячейку. Имя второй переменной окажется в том же столбце под именем первой, то есть во второй строке; имя третьей переменной — в третьей строке, и т. д. Для перемещения между строками пользуйтесь клавишами ↓ и ↑.

Часто шрифт, установленный в SPSS по умолчанию, несовместим с кириллицей. Для того чтобы иметь возможность задавать имена на русском языке, попробуйте поменять шрифт. Для этого в меню View (Вид) выберите команду Fonts (Шрифты) и в открывшемся диалоговом окне Шрифт выберите любой из тех шрифтов, у которых в раскрывающемся списке Набор символов помимо других пунктов есть пункт Кириллица. Выбрав такой шрифт и пункт Кириллица в списке Набор символов, вы получите возможность задавать имена и параметры переменных на русском языке. Если вы пользуетесь одной из ранних версией SPSS, то, возможно, вам и не удастся русифицировать программу. Ничего страшного: задавайте имена и параметры переменных латинскими буквами.

Шаг 3 Для задания имен всем переменным файла ex01.sav выполните следующие действия:

1. Введите символ № и нажмите клавишу ↓.
2. Введите слово пол и нажмите клавишу ↓.
3. Введите слово класс и нажмите клавишу ↓.

Аналогичным образом вводятся оставшиеся имена переменных.

Как вы, наверное, уже обратили внимание, при переходе на следующую строку (то есть при окончании ввода имени переменной) 8 из 9 оставшихся параметров автоматически заполняются значениями по умолчанию.

Имя переменной не является произвольным. Существует ряд соглашений, которым оно должно удовлетворять:

- длина имени не должна превышать 8 символов;
- в имени могут использоваться любые буквы, цифры, символы @, #, ., _, $, однако имя всегда должно начинаться с буквы, а символ . (точка) не может стоять в конце имени;
- имена всех переменных должны быть разными;
- буквы верхнего и нижнего регистров символов не различаются, то есть имена ID, id, Id и iD воспринимаются программой как идентичные;
- имена переменных не должны совпадать с каким-либо из зарезервированных слов (all, ne, eq, to, le, It, by, or, gt, and, not, ge, with).

Следующим шагом является определение типа переменной, что позволит вам максимально эффективно использовать ваши данные в будущем.

## Тип переменной

Параметр Type (Тип) определяет тип переменной. Если фокус ввода находится в столбце Name (Имя) какой-либо из созданных переменных, переместите его в столбец Type (Тип), например, нажав клавишу Tab, тогда в правой части ячейки появится кнопка с многоточием. Щелчок на ней приводит к появлению на экране диалогового окна Variable Type (Тип переменной), представленного на рис. 3.3.

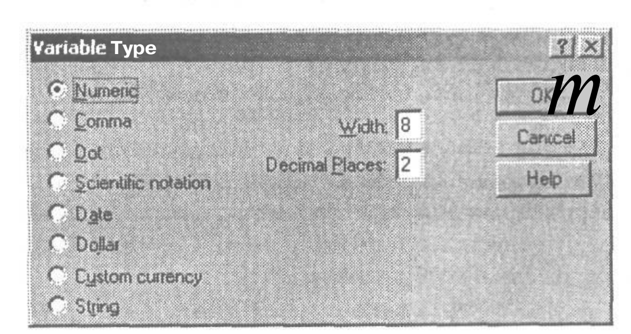



**Рис. 3.3.** Диалоговое окно Variable Type

Как видите, текущим типом переменной является тип Numeric (Числовой). В подавляющем большинстве случаев вам придется иметь дело именно с числовыми данными. В тех редких случаях, когда значения переменных представляют собой буквы или буквосочетания (слова), необходимо установить переключатель String

(Строчный). Строчные данные, в отличие от числовых, могут включать буквы и другие символы, то есть нести текстовую информацию. В частном случае строчная переменная может хранить число, однако обработка такого «числа» будет производиться так, как будто оно является текстом. Как правило, строчные переменные не подлежат обработке. Поэтому их следует избегать, за редким исключением — например, когда данная переменная содержит имена людей.

Необходимость в выборе строчного типа возникает, когда данные переносятся из программы Excel в SPSS путем копирования и вставки. Если значения какой-либо из переменных являются буквенными (например, «м» и «ж»), то перед вставкой этой переменной необходимо изменить ее тип на строчный. В противном случае буквенные значения вставлены не будут.

Остальные 6 типов переменных, которые можно задать с помощью переключателей, присутствующих в диалоговом окне, практически не используются при обработке данных, и поэтому мы не будем их рассматривать.

Поскольку все переменные имеют установленный по умолчанию тип Numeric (Числовой), никаких дополнительных действий с ними производить не нужно.

## Дробная часть числа

Параметр Decimals (Дробная часть) предназначен для задания числа десятичных знаков после запятой в случае, если тип переменной допускает использование дробных чисел. Для строковых переменных значение в ячейке Decimals (Дробная часть) автоматически устанавливается равным нулю, а для цифровых переменных — равным 2. У строковых переменных значение параметра Decimals (Дробная часть) недоступно для изменения.

Когда фокус ввода оказывается в ячейке столбца Decimals (Дробная часть), справа появляются две кнопки счетчика. Щелкая на этих кнопках, вы можете изменять текущее значение ширины переменной. Как вы можете видеть, изначально для всех переменных по умолчанию отведено 2 знака.

В файле ex01.sav наличие знаков после запятой необходимо для переменных отметка1 и отметка2, причем число этих знаков равно 2 и установлено программой по умолчанию. Таким образом, задача заключается в том, чтобы установить для всех остальных цифровых переменных значение параметра Decimals (Дробная часть) равным 0. Отметим, что это не является необходимым, а служит лишь для удобства отображения данных на экране или в отчете.

## Ширина переменной

Параметр Width (Ширина) позволяет задать максимальное количество знаков, которое может иметь значение переменной, включая дробную часть. Например, ширина переменной № (идентификатор) составляет 3 знака, поскольку все числа, которые мы намерены вводить в нее, содержат от 1 до 3 знаков. На практике

определить заранее ширину переменной гораздо труднее, поскольку не всегда известно, какие данные нам будет нужно вводить в нее в будущем. Поэтому следует задавать ширину переменной с гарантированным запасом: ее можно ограничить потом, после ввода данных.

Шаг 4 Для задания числа десятичных знаков в переменных выполните следующие действия:

1. С помощью клавиш со стрелками переместите фокус ввода в ячейку столбца Decimals (Дробная часть), соответствующую переменной №, и дважды щелкните на нижней кнопке счетчика. В ячейке должно остаться значение, равное 0.
2. С помощью клавиш со стрелками переместите фокус ввода в ячейку столбца Decimals (Дробная часть), соответствующую переменной пол, и дважды щелкните на нижней кнопке счетчика. В ячейке должно остаться значение, равное 0.
3. Аналогичные действия произведите с остальными переменными, кроме переменных отметка1 и отметка2.

Для переменных отметка1 и отметка2 (средние отметки учащихся) мы отведем три знака: один знак — это целая часть числа, и два знака после запятой. Переменная пол требует лишь одного знака, поскольку этого достаточно, чтобы определить пол студента.

Шаг 5 Для задания ширины переменных выполните следующие действия:

1. С помощью клавиш со стрелками переместите фокус ввода в ячейку столбца Width (Ширина), соответствующую переменной №, и задайте значение 3. В ячейке должно остаться значение ширины, равное 3.
2. С помощью клавиш со стрелками переместите фокус ввода в ячейку столбца Width (Ширина), соответствующую переменной пол, и задайте значение 1. В ячейке должно остаться значение ширины, равное 1.
3. Перемещая фокус ввода в ячейки столбца Width (Ширина) остальных переменных, задайте ширину каждой переменной в соответствии со значениями переменных в файле ex01.sav.

## Метки переменных

С помощью параметра Label (Метка) можно создать метку переменной. Как правило, метка используется в тех случаях, когда содержание переменной недостаточно ясно отражено в ее названии. По сути метка — это комментарий к имени переменной. При желании вы можете отобразить метки переменных в окне вывода рядом с их именами.

Ячейки столбца Label (Метка) представляют собой обычные текстовые поля, в которые вы можете вводить текст меток. Длина метки не должна превышать 256 символов. Ограничений на используемые символы нет. Но помните о том, что слишком длинные метки ухудшают удобочитаемость фрагментов, в которых

они присутствуют. Как показывает практика, 20-30 символов вполне достаточно, чтобы описать назначение переменной. И чем короче метка, тем лучше.

Если какая-либо из надписей окажется длиннее текущей ширины столбца, последняя автоматически увеличится до необходимого размера. Если вам не обязательно видеть все надписи целиком, вы можете скорректировать размер столбца вручную. Для этого подведите указатель мыши к правой границе ячейки с меткой столбца — указатель примет вид двух стрелок, направленных в разные стороны. Нажмите левую кнопку мыши и, не отпуская ее, перетащите правую границу столбца на новое место. Аналогичным способом можно перетаскивать границу любых столбцов и тем самым оформлять вкладку Variable View (Просмотр переменных) по своему вкусу.

Шаг 6 Для ввода меток выполните следующие действия:

1. С помощью клавиш со стрелками переместите фокус ввода в ячейку столбца Label (Метка), соответствующую переменной тест1, введите словосочетание счет в уме и нажмите клавишу Enter.
2. С помощью клавиш со стрелками переместите фокус ввода в ячейку столбца Label (Метка), соответствующую переменной тест2, введите словосочетание числовые ряды и нажмите клавишу Enter.
3. Для переменной тест3 введите слово словарь, для переменной тест4 — слово осведомленность, для переменной тест5 — словосочетание кратковременная память.

# Метки значений переменных

Параметр Values (Значения) позволяет управлять наименованиями уровней (категорий) переменной. Под уровнем, или категорией, понимается целочисленное значение переменной, имеющее определенный смысл. Например, переменная пол имеет два уровня: 1 — жен (женский) и 2 — муж (мужской). Буквосочетания или слова, поставленные в соответствие уровням переменной, например, жен и муж, называются *метками значений* и отражают смысл разных значений переменной для исследователя. Наличие меток значений исключительно важно для удобочитаемости результатов анализа, поскольку очень трудно запомнить смысл значений всех переменных, даже если их число сравнительно невелико. В крупных демографических исследованиях число переменных может превышать 100, поэтому метки значений являются практически незаменимым средством расшифровки результатов статистического анализа. SPSS также позволяет включать метки значений в файл данных, что очень удобно при создании последнего. Длина метки значения ограничивается 60 символами, однако, как правило, для пользователей достаточно 3-10 символов.

Как и в случае столбца Type (Тип), перемещение в ячейку столбца Values (Значения) приводит к появлению кнопки с многоточием. Если щелкнуть на ней, на экране появится диалоговое окно Value Labels (Метки значений), показанное на рис. 3.4, в котором можно задать метки для различных значений переменной.

**Рис. 3.4.** Диалоговое окно Value Labels

Шаг 7 Ниже в качестве примера приведены действия, позволяющие задать метки значений для переменной пол.

1. С помощью клавиш со стрелками переместите фокус ввода в ячейку столбца Values (Значения), соответствующую переменной пол, и щелкните мышью на кнопке с многоточием.
2. В поле Value (Значение) введите число 1, нажмите клавишу Tab для перехода в поле Value Label (Метка значения), введите туда слово жен и щелкните па кнопке Add (Добавление) или нажмите клавиши Alt+A.
3. В поле Value (Значение) введите число 2, нажмите клавишу Tab для перехода в поле Value Label (Метка значения), введите туда слово муж и щелкните на кнопке Add (Добавление) или нажмите клавиши Alt+A.
4. Для закрытия диалогового окна Value Labels (Метки значений) щелкните на кнопке OK.

После задания меток значений переменных появляется еще одна удобная возможность изменения вида окна Data View (Просмотр данных). Если в меню View (Вид) установить флажок рядом с командой Value Labels (Метки значений), то численные значения переменных будут заменены заданными метками. Для того чтобы вернуться к численным значениям переменных, достаточно сбросить флажок рядом с командой Value Labels (Метки значений).

## Пропуски

Параметр Missing (Пропуски) используется очень редко, поскольку программа и так позволяет учитывать пропуски в данных. Необходимость в этом параметре возникает, когда требуется различать причины пропусков значений. Например, пропуск в данных может быть обусловлен тем, что респондент еще не опрошен, а может быть и так, что он отказался отвечать на данный вопрос. Так, в отношении переменной вуз (предполагаемый для поступления вуз) нам необходимо различать тех учащихся, которых мы не успели опросить, и тех, которые еще не определились. Для еще не опрошенных учащихся мы будем оставлять пустую ячейку, а неопределившихся учащихся будем обозначать цифрой 9. Если ввести значение 9 в столбец Missing (Пропуски), то оно не будет использоваться в дальнейшем при обработке наряду с пустыми ячейками. После щелчка на

кнопке с многоточием, появляющейся при перемещении фокуса ввода в ячейку столбца Missing (Пропуски), открывается диалоговое окно Missing Values (Отсутствующие данные), в котором можно определить необходимые значения.

## Столбцы

Параметр Columns (Столбцы), в отличие от своего соседа слева, требуется для всех переменных. С его помощью можно управлять шириной (в символах) столбцов вкладки просмотра данных (Data View). Если вы зададите широкие столбцы, то это позволит полностью видеть имена переменных, однако число одновременно видимых переменных будет невелико. Напротив, узкие столбцы позволяют наблюдать на экране много переменных сразу, но их имена могут не уместиться в столбце целиком. Например, при ширине столбцов в 3 символа на экране можно увидеть 33 переменные, а при ширине в 8 символов — всего 12. Опытные пользователи знают, что работа с данными тем удобнее, чем больше их видно одновременно. Поэтому вы часто будете сталкиваться с задачей размещения на экране как можно большего объема данных с сохранением их удобочитаемости. Пожалуй, основной принцип, которым следует руководствоваться — чем меньше знаков занимает большая часть значений некой переменной, тем меньшей следует задавать ширину ее столбца.

Для изменения параметра Columns (Столбцы) используется счетчик. По умолчанию для всех переменных значение Columns (Столбцы) равно 8.



Шаг 8 Для примера ниже приведены действия по изменению ширины столбцов для переменных №, пол, класс.

1. С помощью клавиш со стрелками переместите фокус ввода в ячейку столбца Columns (Столбцы), соответствующую переменной №, и несколько раз щелкните на нижней кнопке счетчика, пока не останется значение 4.
2. С помощью клавиш со стрелками переместите фокус ввода в ячейку столбца Columns (Столбцы), соответствующую переменной пол, и несколько раз щелкните на нижней кнопке счетчика, пока не останется значение 6.
3. С помощью клавиш со стрелками переместите фокус ввода в ячейку столбца Columns (Столбцы), соответствующую переменной класс, и три раза щелкните на нижней кнопке счетчика. В ячейке должно остаться значение 5. Подобным образом подберите ширину оставшихся столбцов, проверяя результат на вкладке Data View (Просмотр данных).

## Выравнивание

Параметр Align (Выравнивание) позволяет управлять расположением данных внутри ячейки. С помощью раскрывающегося списка, доступ к которому появляется при получении ячейкой фокуса ввода, можно выбрать один из трех типов выравнивания: Left (Влево), Right (Вправо) и Center (По центру). По умолчанию значения цифровых переменных выравниваются по правому краю, строчных — по левому. Для каждой переменной вы можете самостоятельно задать тип выравнивания.

## Шкала измерения

Как и у предыдущего параметра, значение параметра Measure (Шкала измерения) выбирается в раскрывающемся списке из трех доступных: Scale (Метрическая), Ordinal (Порядковая) и Nominal (Номинативная). Каждое из этих значений несет дополнительную информацию о типе данных переменной.

- Значение Scale (Метрическая) указывается для количественных типов данных, допускающих арифметические операции. Примером может служить возраст человека: для него имеют смысл операции вычитания, сложения и деления. В файле ex01.sav это значение параметра Measure (Шкала измерения) можно указать для переменных test1, ..., test5, отметка1, отметка2.
- Значение Ordinal (Порядковая) указывается для количественных данных, отражающих измеренное качество на уровне порядка и не допускающих основных арифметических операций. Например, упорядочивание учащихся по сообразительности или кодирование ответов по степени выраженности (1 — «не нравится», 2 — «нейтрально», 3 — «правится») — это измерения в порядковой шкале. К подобным типам можно отнести и многие дихотомические переменные (переменные, имеющие лишь два различных значения: 0 или 1, да или нет и т. п.).
- Значение Nominal (Номинативная) указывается для категориальных (неколичественных) типов данных, не отражающих количество измеряемого свойства. В файле ex01 .sav к таким данным относятся переменные №, пол, класс, вуз, хобби.

Иногда может оказаться затруднительным сделать выбор между значениями Scale (Метрическая) и Ordinal (Порядковая). Не отчаивайтесь, если попадете в такую ситуацию: во всех статистических процедурах оба этих значения параметра Measure (Шкала измерения) сказываются на результатах анализа одинаковым образом.

Чаще всего нет необходимости менять заданное по умолчанию значение этого параметра Scale (Метрическая). Тем более что выбор значения Nominal (Номинативная) ограничивает набор возможных методов обработки данных.

Итак, когда все переменные созданы, а их параметры определены, настало время заняться вводом данных.

# Ввод данных

Ввод самих данных гораздо проще, чем задание параметров переменных, поэтому самое трудное в этой главе для вас уже позади. В этом разделе мы опишем два способа непосредственного ввода данных в программу, а также процедуру переноса данных из программы Excel. Но сначала рассмотрим действия, необходимые для сохранения файла.

В процессе ввода рекомендуется время от времени производить сохранение файла во избежание случайной порчи или утери введенных данных. Обратите внимание на то, что перед вводом данных следует перейти на вкладку Data View (Просмотр данных), показанную ранее на рис. 3.1.

**Шаг 9** В случае если вы сохраняете файл в первый раз, потребуется задать его имя.

1. В меню File (Файл) выберите команду Save (Сохранение). На экране появится стандартное диалоговое окно сохранения файла.
2. Введите имя файла ex01.sav и щелкните на кнопке Save (Сохранить).

Обратите внимание на то, что указанные действия позволяют сохранять файлы лишь в текущей папке. Если вы создали рабочий каталог, то текущей является папка C:\SPSSWORK, если нет, то папка C:\Program Files\SPSS. При необходимости вы можете сохранить файл в любой другой папке обычным способом.

**Шаг 10** Если ваш файл уже был сохранен хотя бы один раз, то достаточно выбрать команду File ▸ Save (Файл ▸ Сохранение) или щелкнуть на кнопке Save File (Сохранение файла) панели инструментов.

Существует два метода непосредственного ввода данных в ячейки электронной таблицы: *по строкам (переменным)* и *по столбцам (объектам)*. В зависимости от конкретных данных тот или другой способ оказывается предпочтительным.

Ввод данных *по переменным* предполагает последовательное заполнение всех строк значениями одной переменной, затем другой переменной и т. д. Вы вводите значение в левую верхнюю ячейку окна, затем нажимаете клавишу ↓ или Enter, переходя на следующую строку, заполняете еще одну ячейку и продолжаете эти действия до тех пор, пока не заполните весь первый столбец. Затем заполняете столбец, соответствующий второй переменной, и т. д., пока все переменные не будут введены. Таким образом, с визуальной точки зрения ввод данных по переменным осуществляется «по вертикали».

Ввод данных *по объектам* заключается в последовательном заполнении каждой строки значениями всех переменных. Сначала вводится значение в левую верхнюю ячейку окна, затем с помощью клавиши → или Tab осуществляется переход в соседний столбец (переменную), вводится значение переменной и т. д. до тех пор, пока вся строка не будет заполнена. После этого аналогичным способом заполняются вторая, третья и все остальные строки. В отличие от предыдущего способа ввод данных по объектам осуществляется «по горизонтали».

Некоторым пользователям привычнее вводить, хранить и редактировать данные в электронной таблице программы Excel. Следует отметить, что программа SPSS лучше приспособлена для этого. Тем не менее перенос данных из Excel в SPSS не составляет труда, если в таблице данных программы Excel строки соответствуют объектам, а столбцы — переменным. Сначала необходимо подготовить структуру файла исходных данных в SPSS, как это было описано ранее. Прежде всего обращайте внимание на те переменные, которые имеют буквенные значения. Для них вместо принятого по умолчанию числового типа переменной необходимо задать строчный тип, установив в окне Variable Type (Тип переменной), показанном на рис. 3.3, переключатель String (Строчный). После задания структуры данных, не закрывая подготовленный файл SPSS, откройте файл Excel, содержащий данные. Выделите и скопируйте при помощи мыши необходимый блок данных (только

значения переменных) в таблице Excel (если эта процедура вызывает у вас затруднения, изучите сначала раздел «Редактирование данных»). Перейдите к файлу SPSS и выделите верхнюю правую ячейку пока еще пустой таблицы данных и в меню Edit (Редактирование) выберите команду Paste (Вставка). Таблица SPSS будет заполнена теми данными, которые вы ранее скопировали из таблицы Excel. Сходным образом можно также пополнять данные, перенося их из Excel в SPSS и наоборот.

# Редактирование данных

Иногда у вас может возникнуть необходимость изменить данные, уже введенные в файл. Мы рассмотрим, каким образом скорректировать содержимое ячейки, создать новый объект (строку) или переменную (столбец), а также скопировать, удалить и вставить данные.

## Изменение содержимого ячейки

Щелкните на ячейке, содержимое которой вы намерены отредактировать, введите новое значение, а затем перейдите в любую соседнюю ячейку с помощью клавиш Enter, Tab или клавиш со стрелками.

## Вставка нового объекта

Если вам необходимо вставить новый объект (строку) между двумя соседними строками, щелкните сначала на нижней из них, а затем — на кнопке Insert Cases (Вставка объектов) панели инструментов. В результате будет создана пустая строка, а номера строк, находящихся ниже, увеличатся на единицу.

## Вставка новой переменной

Чтобы вставить новую переменную между двумя соседними, щелкните сначала на правой из них, а затем — на кнопке Insert Variable (Вставка переменной). Будет создан пустой столбец, а все переменные, находящиеся справа, окажутся сдвинутыми на один столбец.

## Копирование и вырезание содержимого ячеек

Для того чтобы скопировать или вырезать содержимое одной или нескольких ячеек, сначала необходимо их выделить. Активная ячейка (то есть ячейка, в которой находится фокус ввода) всегда является выделенной. Чтобы выделить группу ячеек, нажмите левую кнопку мыши на угловой ячейке будущей группы, и удерживая эту кнопку, перетащите указатель к противоположной угловой ячейке группы, после чего кнопку мыши отпустите. Если вам необходимо выделить целиком строку или столбец, щелкните соответственно на нужном объекте или на имени переменной.

Когда желаемые ячейки выделены, в меню Edit (Редактирование) выберите команду Copy (Копирование) или Cut (Вырезание). В обоих случаях содержимое ячеек будет скопировано в буфер обмена, однако команда Copy (Копирование) оставит исходные ячейки нетронутыми, а команда Cut (Вырезание) очистит их.

## Вставкаячеек

Для того чтобы вставить предварительно скопированные в буфер обмена данные, нужно переместить фокус ввода в левую верхнюю ячейку группы, в которую будет осуществляться вставка, и в меню Edit (Редактирование) выбрать команду Paste (Вставка). Следует помнить о двух неприятностях, связанных с операцией вставки.

- Если в области, в которую планируется вставить данные, уже содержатся какие-либо другие данные, то после выполнения вставки они будут утеряны.
- При копировании и вставке данных их смысл не анализируется программой, и вы можете случайно заменить значения одной переменной значениями любой другой переменной. Последствия такой вставки могут быть весьма плачевными, поскольку разрешить возникшую путаницу, как правило, весьма непросто.

Мы приводим несколько рекомендаций, которые могут быть полезными при выполнении вставки.

- Перед удалением и вставкой важных фрагментов данных сохраняйте текущий файл. В случае ошибки вы сможете открыть этот файл снова.
- По возможности создавайте новые переменные и строки для вставки данных.
- Как правило, вам придется иметь дело с копированием, вырезанием и вставкой столбцов (переменных) или строк (объектов) целиком. Однако особую аккуратность следует проявлять при вставке *блоков* столбцов или строк.
- Будьте особенно внимательны при вставке значений переменных разных типов, например, не допускайте вставки значений строчных переменных в столбцы, соответствующие числовым переменным. Это может привести к трудно идентифицируемым ошибкам в процессе статистического анализа.

## Поискданных

Очень удобным вспомогательным средством работы с данными является функция поиска. В меню Edit (Редактирование) выберите команду Find (Поиск) или щелкните на кнопке Find (Поиск) панели инструментов. На экране появится диалоговое окно Search Data in Variable <name> (Поиск данных в переменной <имя>), с помощью которого можно найти заданное слово или значение (рис. 3.5). Как правило, команда Find (Поиск) применяется для поиска имен в больших файлах данных, а также для поиска значений, которые требуется изменить.

В случае если вы, распечатав итоговые данные, обнаружите недопустимые или неверные значения какой-либо из переменных, можете применить приведенный алгоритм для поиска этого значения и исправления ошибки.

**Рис. 3.5.** Диалоговое окно поиска

**Шаг 11** Попробуем, к примеру, найти в файле ex01.sav ученика с номером 93.

1. На вкладке Data View (Просмотр данных) щелкните на имени переменной №, затем в меню Edit (Редактирование) выберите команду Find (Поиск) или щелкните на кнопке Find (Поиск) панели инструментов.
2. В поле Find what (Найти) открывшегося диалогового окна Search Data in Variable № (Поиск данных в переменной №) введите число 93. Если при поиске в строчной переменной требуется различать строчные и прописные буквы, установите флажок Match case (Учитывать регистр). Щелкните на кнопке Find next (Найти далее), чтобы найти ученика с номером 93.

После завершения поиска щелкните на кнопке Cancel (Отмена), чтобы закрыть диалоговое окно.

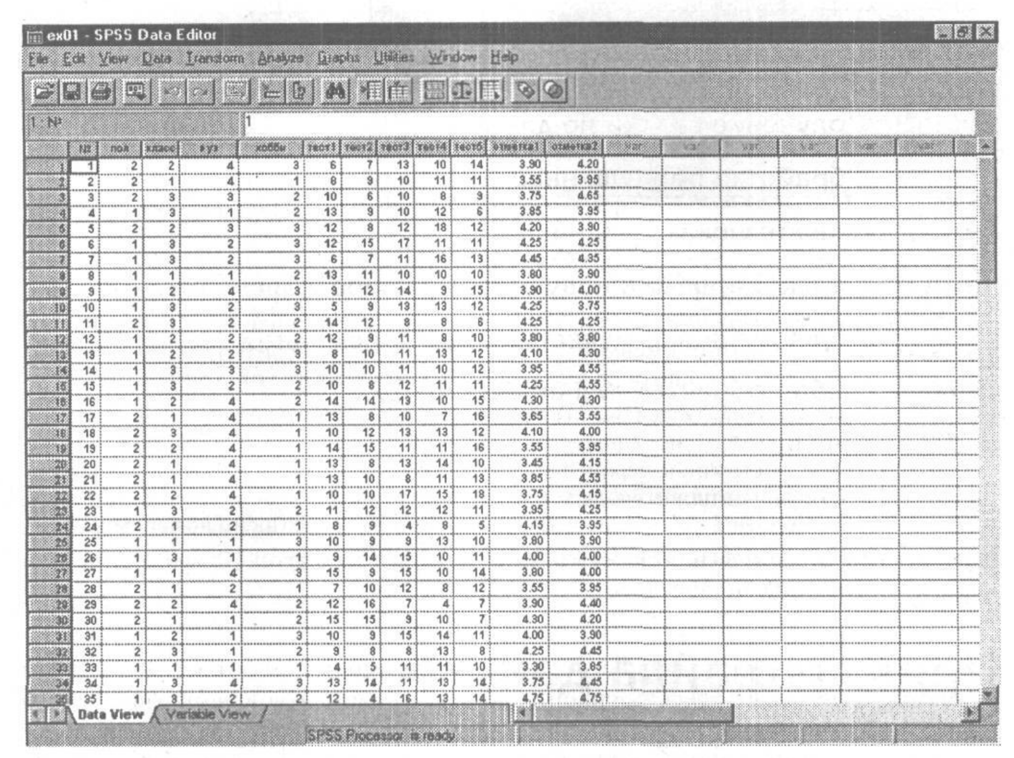
# Пример файла данных

Файл данных ex01.sav содержит информацию об учащихся трех классов. Переменные для этого файла перечислены в табл. 3.2.

**Таблица 3.2.** Переменные файла данных exOl .sav

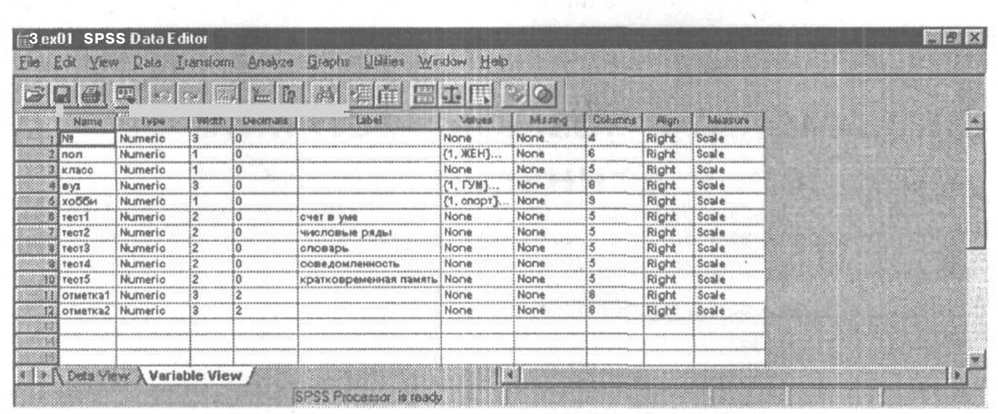
| Переменная | Описание |
|---|---|
| № | Шестизначный идентификационный номер студента |
| пол | [illegible] |
| класс | Класс, в котором учится школьник (1 — «А», 2 — «Б». 3 — «В») |
| вуз | Предполагаемый для поступления вуз: (1 — гуманитарный, 2 — экономический, 3 — технический, 4 — естественнонаучный) |
| хобби | Внешкольное увлечение ученика (1 — спорт, 2 — компьютеры, 3 — искусство) |
| тест1 | Показатель теста «Счет в уме» |
| тест2 | Показатель теста «Числовые ряды» |
| тест3 | Показатель теста «Словарный запас» |
| тест4 | Показатель теста «Осведомленность» |
| тест5 | Показатель теста «Кратковременная вербальная память» |
| отметка1 | Средний балл отметок за 10-й класс |
| отметка2 | Средний балл отметок за 11-й класс |

Вкладки Data View (Просмотр данных) и Variable View (Просмотр переменных) окна редактора данных для файла ex01 .sav показаны соответственно на рис. 3.6 и 3.7.



|  | № | пол | класс | вуз | хобби | тест1 | тест2 | тест3 | тест4 | тест5 | отметка1 | отметка2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 2 | 2 | 4 | 3 | 6 | 7 | 13 | 10 | 14 | 3.90 | 4.20 |
| 2 | 2 | 2 | 1 | 4 | 1 | 8 | 9 | 10 | 11 | 11 | 3.55 | 3.95 |
| 3 | 3 | 2 | 3 | 3 | 2 | 10 | 6 | 10 | 8 | 9 | 3.75 | 4.65 |
| 4 | 4 | 1 | 3 | 1 | 2 | 13 | 9 | 10 | 12 | 6 | 3.85 | 3.95 |
| 5 | 5 | 2 | 2 | 3 | 3 | 12 | 8 | 12 | 18 | 12 | 4.20 | 3.90 |
| 6 | 6 | 1 | 3 | 2 | 3 | 12 | 15 | 17 | 11 | 11 | 4.25 | 4.25 |
| 7 | 7 | 1 | 3 | 2 | 3 | 6 | 7 | 11 | 16 | 13 | 4.45 | 4.35 |
| 8 | 8 | 1 | 1 | 1 | 2 | 13 | 11 | 10 | 10 | 10 | 3.80 | 3.90 |
| 9 | 9 | 1 | 2 | 4 | 3 | 9 | 12 | 14 | 9 | 15 | 3.90 | 4.00 |
| 10 | 10 | 1 | 3 | 2 | 3 | 5 | 9 | 13 | 13 | 12 | 4.25 | 3.75 |
| 11 | 11 | 2 | 3 | 2 | 2 | 14 | 12 | 8 | 8 | 6 | 4.25 | 4.25 |
| 12 | 12 | 1 | 2 | 2 | 2 | 12 | 9 | 11 | 8 | 10 | 3.80 | 3.80 |
| 13 | 13 | 1 | 2 | 2 | 3 | 8 | 10 | 11 | 13 | 12 | 4.10 | 4.30 |
| 14 | 14 | 1 | 3 | 3 | 3 | 10 | 10 | 11 | 10 | 12 | 3.95 | 4.55 |
| 15 | 15 | 1 | 3 | 2 | 2 | 10 | 8 | 12 | 11 | 11 | 4.25 | 4.55 |
| 16 | 16 | 1 | 2 | 4 | 2 | 14 | 14 | 13 | 10 | 15 | 4.30 | 4.30 |
| 17 | 17 | 2 | 1 | 4 | 1 | 13 | 8 | 10 | 7 | 16 | 3.65 | 3.55 |
| 18 | 18 | 2 | 3 | 4 | 1 | 10 | 12 | 13 | 13 | 12 | 4.10 | 4.00 |
| 19 | 19 | 2 | 2 | 4 | 1 | 14 | 15 | 11 | 11 | 16 | 3.55 | 3.95 |
| 20 | 20 | 2 | 1 | 4 | 1 | 13 | 8 | 13 | 14 | 10 | 3.45 | 4.15 |
| 21 | 21 | 2 | 1 | 4 | 1 | 13 | 10 | 8 | 11 | 13 | 3.85 | 4.55 |
| 22 | 22 | 2 | 2 | 4 | 1 | 10 | 10 | 17 | 15 | 18 | 3.75 | 4.15 |
| 23 | 23 | 1 | 3 | 2 | 2 | 11 | 12 | 12 | 12 | 11 | 3.95 | 4.25 |
| 24 | 24 | 2 | 1 | 2 | 1 | 8 | 9 | 4 | 8 | 5 | 4.15 | 3.95 |
| 25 | 25 | 1 | 1 | 1 | 3 | 10 | 9 | 9 | 13 | 10 | 3.80 | 3.90 |
| 26 | 26 | 1 | 3 | 1 | 1 | 9 | 14 | 15 | 10 | 11 | 4.00 | 4.00 |
| 27 | 27 | 1 | 1 | 4 | 3 | 15 | 9 | 15 | 10 | 14 | 3.80 | 4.00 |
| 28 | 28 | 2 | 1 | 2 | 1 | 7 | 10 | 12 | 8 | 12 | 3.55 | 3.95 |
| 29 | 29 | 2 | 2 | 4 | 2 | 12 | 16 | 7 | 4 | 7 | 3.90 | 4.40 |
| 30 | 30 | 2 | 1 | 1 | 2 | 15 | 15 | 9 | 10 | 7 | 4.30 | 4.20 |
| 31 | 31 | 1 | 2 | 1 | 3 | 10 | 9 | 15 | 14 | 11 | 4.00 | 3.90 |
| 32 | 32 | 2 | 3 | 1 | 2 | 9 | 8 | 8 | 13 | 8 | 4.25 | 4.45 |
| 33 | 33 | 1 | 1 | 1 | 1 | 4 | 5 | 11 | 11 | 10 | 3.30 | 3.85 |
| 34 | 34 | 1 | 3 | 4 | 3 | 13 | 14 | 11 | 13 | 14 | 3.75 | 4.45 |
| 35 | 35 | 1 | 3 | 2 | 2 | 12 | 4 | 16 | 13 | 14 | 4.75 | 4.75 |

**Рис. 3.6.** Данные файла ex01.sav



|  | Name | Type | Width | Decimals | Label | Values | Missing | Columns | Align | Measure |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | № | Numeric | 3 | 0 |  | None | None | 4 | Right | Scale |
| 2 | пол | Numeric | 1 | 0 |  | {1, ЖЕН}... | None | 6 | Right | Scale |
| 3 | класс | Numeric | 1 | 0 |  | None | None | 5 | Right | Scale |
| 4 | вуз | Numeric | 3 | 0 |  | {1, ГУМ}... | None | 8 | Right | Scale |
| 5 | хобби | Numeric | 1 | 0 |  | {1, спорт}... | None | 9 | Right | Scale |
| 6 | тест1 | Numeric | 2 | 0 | счет в уме | None | None | 5 | Right | Scale |
| 7 | тест2 | Numeric | 2 | 0 | числовые ряды | None | None | 5 | Right | Scale |
| 8 | тест3 | Numeric | 2 | 0 | словарь | None | None | 5 | Right | Scale |
| 9 | тест4 | Numeric | 2 | 0 | осведомленность | None | None | 5 | Right | Scale |
| 10 | тест5 | Numeric | 2 | 0 | кратковременная память | None | None | 5 | Right | Scale |
| 11 | отметка1 | Numeric | 3 | 2 |  | None | None | 8 | Right | Scale |
| 12 | отметка2 | Numeric | 3 | 2 |  | None | None | 8 | Right | Scale |

**Рис. 3.7.** Переменные файла ex01.sav

# 4 Управление данными



Материал этой главы посвящен вопросам эффективной работы с исходными данными. Описанные здесь операции весьма полезны в большинстве случаев обработки и анализа данных, так как практически всегда существует необходимость в предварительной подготовке и преобразовании исходных данных. Поэтому изложенные рекомендации по форматированию данных помогут вам работать с программой гораздо свободнее.

## Знакомство с возможностями управления данными

В процессе работы вам могут понадобиться *агрегированные* данные, то есть данные, являющиеся результатом некоторых действий над исходными данными файла. К примеру, для вашего исследования может представлять интерес среднее значение или сумма баллов по нескольким тестам для каждого учащегося, ранг каждого ученика по успеваемости и т. п. Иногда желательно упорядочить данные файла по какому-либо признаку, например по результатам выполнения какого-либо задания. Нередко возникает необходимость обработки не всех данных файла, а лишь их подмножества, выделяемого по определенным критериям (например, по полу, классу, успеваемости и пр.). Существует и обратная задача: если данные хранятся в нескольких небольших файлах, может возникнуть потребность в их объединении для последующего анализа.

Перечисленные проблемы указывают на то, что для регулярной аналитической работы недостаточно умения вводить данные и применять к ним статистические процедуры. Возникает задача эффективного управления данными. Способы решения этой задачи бывают весьма нетривиальными, и исчерпывающий рассказ о них не вполне соответствовал бы теме книги. Тем не менее представленного в этой главе материала вполне достаточно, чтобы научиться свободно манипулировать данными даже новичку.

Несмотря на то что навыки управления данными приходят с опытом и требуют некоторого терпения, обязательно освойте их. Это придаст процессу исследования гибкость, простоту и легкость. Тогда выполнение статистических процедур, казавшихся сложными и громоздкими, станет для вас интуитивно понятным.

Мы рассмотрим следующие основные команды управления данными:

- команда Analyze ▸ Reports ▸ Case Summaries (Анализ ▸ Отчеты ▸ Сводка по данным) предназначена для проверки состава и качества данных;
- команда Transform ▸ Replace Missing Values (Преобразование ▸ Заменить пропущенные значения), как ясно из ее названия, работает с отсутствующими значениями переменных;
- команда Transform ▸ Compute (Преобразование ▸ Вычислить) позволяет путем вычислений создавать новые переменные на основе существующих;
- команда Transform ▸ Rank Cases (Преобразование ▸ Ранжировать объекты) позволяет создать новую переменную путем ранжирования значений существующей переменной;
- команды подменю Recode (Перекодировка) меню Transform (Преобразование) предназначены для изменения способа кодирования переменных, например уменьшения числа возможных значений;
- с помощью команды Data ▸ Select Cases (Данные ▸ Выбор объектов) можно выбрать подмножество объектов для дальнейшего анализа;
- команда Data ▸ Sort Cases (Данные ▸ Сортировка объектов) позволяет упорядочить объекты в соответствии с назначенными критериями;
- команды подменю Merge Files (Слияние файлов) меню Data (Данные) используются для добавления в файл новых переменных или объектов из другого файла.

Все главы начиная с этой построены таким образом, что практически не зависят друг от друга по содержанию. После вводной части каждой главы идет описание пошаговых процедур, которые позволяют открывать файлы с данными и получать доступ к диалоговым окнам нужных статистических операций. Затем следуют описания этих диалоговых окон, а заключительные части глав посвящены интерпретации выводимых данных и их распечатке. Как правило, действия, связанные с вызовом статистической операции, приводятся на шаге 4, а на шаге 5 рассказывается о том, как задать параметры операции и завершить анализ данных. Иногда имеется несколько альтернативных способов выполнения шага 5, в этом случае они помечаются как 5, 5а, 5б и т. д.

Структура данной главы немного отличается от описанной, поскольку в ней идет речь не о статистических операциях, а о средствах организации данных. После выполнения трех подготовительных шагов будет детально рассмотрена работа с каждой из семи команд управления данными. Главу завершит обзор необходимых завершающих действий с программой. В отличие от глав, посвященных анализу данных, здесь вы не найдете заключительного раздела, относящегося к выводу результатов.

Прежде чем двигаться дальше, выполните три подготовительных шага.

**Шаг 1** Подготовьте новый или существующий файл данных. Все необходимые для этого действия описаны в главе 3.

**Шаг 2** Запустите программу SPSS при помощи значка на рабочем столе или выбрав в главном меню Windows команду Пуск ▸ Программы ▸ SPSS for Windows ▸ SPSS 11.5 for Windows (Start ▸ Programs ▸ SPSS for Windows ▸ SPSS 11.5 for Windows) и в открывшемся после запуска программы диалоговом окне SPSS for Windows щелкните па кнопке Cancel (Отмена).

После выполнения этого шага на экране появится окно редактора данных SPSS.

**Шаг 3** Откройте файл данных, с которым вы намерены работать (в нашем случае это файл ex01.sav). Если он расположен в текущей папке, то выполните следующие действия:

1. Выберите в меню File (Файл) команду Open ▸ Data (Открытие ▸ Данные) или щелкните па кнопке Open File (Открытие файла) панели инструментов.
2. В открывшемся диалоговом окне дважды щелкните на имени ex01.sav или введите его с клавиатуры и щелкните на кнопке OK.

Независимо от того, открыта программа SPSS только что или какие-то процедуры уже выполнялись, в верхней части главного окна должна присутствовать строка меню (она показана ниже). Пока строка меню присутствует на экране, доступны все команды анализа данных. При этом окно с данными видеть не обязательно.

File Edit View Data Transform Analyze Graphs Utilities Window Help

При работе со сводными таблицами или при редактировании диаграмм некоторые пункты строки меню могут исчезать или меняться. Чтобы вернуться к главному окну и стандартной строке меню, щелкните на кнопке свертывания или восстановления текущего окна.

# Получение сводки по данным

Команда Case Summaries (Сводка по данным) предназначена для получения с различной степенью детализации упорядоченного списка всех данных файла или их подмножества. Это обычно необходимо при определении состава и качества данных для их дальнейшего редактирования или анализа.

После завершения шага 3 нужная строка меню должна присутствовать на экране.

Шаг 4 В меню Analyze (Анализ) выберите команду Reports ▸ Case Summaries (Отчеты ▸ Сводка по данным). На экране появится диалоговое окно Summarize Cases (Сводка по данным), представленное на рис. 4.1.

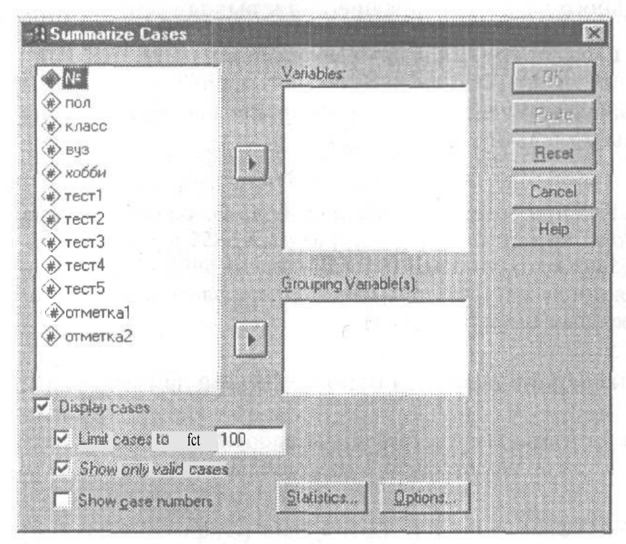



Рис. 4.1. Диалоговое окно Summarize Cases

Список в левой части окна позволяет выбирать переменные для сводки. Назначение остальных элементов окна описано далее.

- В список Variables (Переменные) вы можете перенести все или некоторые переменные для сводки.
- Список Grouping Variable(s) (Группирующая переменная) определяет порядок перечисления объектов. Если список пуст, объекты будут перечислены в том порядке, в котором они перечислены в файле. Если же, например, включить в список Grouping Variable(s) (Группирующая переменная) переменную пол, то сначала будут перечислены все девочки, а затем все мальчики в порядке их положения в файле. Девочки окажутся перечисленными первыми, поскольку значение переменной пол для них равно 1, а для мальчиков — 2. Аналогично, если в список ввести переменную класс, то данные файла будут упорядочены по номеру группы: сначала будут перечислены школьники класса 1 («А»), затем школьники класса 2 («Б») и, наконец, школьники класса 3 («В»). Если в список Grouping Variable(s) (Группирующая переменная) включить несколько переменных, то данные будут упорядочены по значениям первой переменной, затем — по значениям второй переменной, и т. д.
- Флажок Display Cases (Отображать значения) управляет отображением значений переменных, и по умолчанию он установлен. Если его сбросить, результат будет содержать только число значений каждой переменной в различных категориях.

- Флажок Limit cases to first (Ограничить число объектов первыми) позволяет ограничить число объектов (строк) для сводки. Если он установлен, то становится доступным поле, расположенное справа от флажка, в которое можно ввести нужное для учета число объектов. По умолчанию это число равно 100. Если флажок сброшен, то учитываются все объекты.
- Флажок Show only valid cases (Показывать только существующие значения) определяет, нужно ли включать в результат отсутствующие значения переменных. По умолчанию он установлен, однако вы будете часто его сбрасывать, чтобы иметь возможность отследить, какие значения отсутствуют.
- Флажок Show case numbers (Показывать номера объектов) определяет, нужно ли включать в выводимые данные номера объектов. Несмотря на то что по умолчанию он сброшен, практически всегда его необходимо устанавливать, поскольку для корректировки данных вам понадобится знать номер нужного объекта.
- Кнопка Statistics (Статистики) обеспечивает доступ к диалоговому окну, в котором можно построить список описательных статистик (описательные статистики подробно рассматривается в главе 7). По умолчанию выбран тип Number of Cases (Число объектов). Другие типы могут быть перенесены в список Cell Statistics (Статистики для ячеек) тем же способом, что и переменные рассматриваемого нами диалогового окна.
- Кнопка Options (Параметры) позволяет открыть диалоговое окно, в котором можно задать, нужно ли учитывать объекты, содержащие пропущенные значения, а также отредактировать содержимое заголовков и подзаголовков.

**Шаг 5** Ниже приведены действия, приводящие к отображению значений всех переменных для первых 20 объектов, при этом номера объектов включаются в результат (вывод). Предполагаетея, что после выполнения инструкций
1 шага 4 окно Summarize Cases (Сводка по данным) открыто.

1. Нажмите кнопку мыши па верхней переменной в списке; не отпуская кнопки, перетащите указатель вниз на последнюю переменную и отпустите кнопку. В результате окажутся выделенными все переменные.
2. Щелкните па верхней кнопке со стрелкой. Выделенные переменные переместятся из общего списка в список Variables (Переменные).
3. Переведите фокус ввода в поле справа от флажка Limit cases to first (Ограничить число объектов первыми), трижды нажав клавишу Tab (вместо клавиши Tab можно воспользоваться мышью). При этом текущее значение поля окажется выделенным.
4. Введите в поле Limit cases to first (Ограничить число объектов первыми) значение 20, установите флажок Show case numbers (Показывать номера объектов) и щелкните на кнопке OK.

В этом простом примере мы всего лишь ограничили вывод результатов частью данных файла. Вернем окну Summarize Cases (Сводка по данным) исходный вид (см. рис. 4.1) и посмотрим, как выполняется группировка данных.

**Шаг 5а** Чтобы для всех объектов вывести значения переменных №, вуз и хобби, сгруппированные по значениям переменных класс и пол, выполните следующие действия:

1. Удерживая нажатой клавишу Ctrl, щелкните поочередно на переменных №, вуз и хобби, выделив их, а затем щелкните на верхней кнопке со стрелкой. Выделенные переменные переместятся из общего списка в список Variables (Переменные).
2. Выделите щелчком переменную класс и щелкните на нижней кнопке со стрелкой. Переменная класс окажется в списке Grouping Variable(s) (Группирующая переменная). Сделайте то же самое для переменной пол.
3. Дважды нажмите клавишу Tab (вместо клавиши Tab можно воспользоваться мышью), чтобы переместить фокус ввода в поле справа от флажка Limit cases to first (Ограничить число объектов первыми), введите в это поле значение 100, установите флажок Show case numbers (Показывать номера объектов) и щелкните на кнопке OK.

В результате выполнения этой процедуры результаты сводки по данным окажутся упорядоченными. Сначала будут перечислены девочки, обучающиеся в классе «А», затем мальчики из класса «А», затем девочки из класса «Б», и т. д. Для каждого школьника будет выведен его порядковый номер.

# Обработка пропущенных значений

В процессе работы с программой SPSS вы нередко будете сталкиваться с проблемой отсутствующих данных. Обратимся к переменным из примера ex01.sav. Вполне вероятна ситуация, когда кто-либо из учеников отсутствовал при проведении тестирования или не ответил на вопрос о внешкольном увлечении, либо не определился с перспективой поступления в вуз. Подобные случаи приводят к тому, что в данных рабочего файла появляются *пропущенные значения*. Пропущенные значения не только мешают осмысливать данные, но и могут оказывать нежелательное влияние па результаты анализа. Большинство статистических операций игнорируют объекты (строки), в которых содержится хотя бы одно пропущенное значение. Если, к примеру, из 35 объектов 13 имеют пропущенные значения по разным переменным, то анализу будет подлежать немногим более 60 % данных файла, что, несомненно, исказит результаты.

Большинство статистических методов SPSS позволяет учитывать пропуски в данных двумя принципиально различными способами: *построчно* (listwise) и *попарно* (pairwise). При построчном учете пропусков SPSS перед выполнением операции проверяет строки (объекты) на наличие пропущенных значений и в случае обнаружения последних исключает соответствующие строки из анализа целиком. Этот способ позволяет получить наиболее корректные статистические результаты, однако потери данных при этом максимальны. При попарном учете пропусков обработка выполняется без дополнительных проверок, и в процессе вычислений не выполняются только те операции, которые требуют наличия пропущенного

значения. Таким образом, в анализе участвуют все введенные данные, но результаты анализа содержат погрешности.

Мы рекомендуем вам по возможности решать проблему пропущенных значений на этапе ввода и кодирования данных, а не полагаться на то, что SPSS сделает это за вас. В любом случае, чем больше пропусков в исходных данных, тем менее точны и корректны результаты анализа.

Для номинативной переменной проблема пропущенных значений решается легко: вы можете просто ввести еще одну ее градацию, которая соответствует пропуску в данных. Для количественной переменной (метрической или порядковой), имеющей множество возможных значений, в SPSS предусмотрены специальные процедуры для заполнения пропусков: в меню Transform (Преобразование) есть команда Replace Missing Values (Заменить пропущенные значения). При всем соблазне ее применения следует помнить, что результаты обработки данных с заменой пропусков фиктивными значениями, например средними, вряд ли могут вызвать доверие. Поэтому лучше на месте пропуска честно оставлять пустую ячейку. А вопрос о построчном или попарном учете пропусков решать отдельно для каждого конкретного метода анализа данных.

В справочной системе SPSS часто используется два термина: ***физически пропущенные значения*** (system missing values) и ***логически пропущенные значения*** (user missing values). Под физически пропущенными значениями понимаются значения, не введенные в компьютер. В редакторе данных пустые ячейки, не содержащие значения, помечены точкой. Логически пропущенные значения — это специальные значения переменной, отражающие невозможность адекватного кодирования некоторой ситуации. Если, например, 1, 2 и 3 — тестовые оценки испытуемого, 8 означает, что тест не завершен, а 9 фиксирует неявку испытуемого, то значения 8 и 9 относятся к логически пропущенным, поскольку их нельзя интерпретировать как результаты теста.

# Преобразование данных

## Вычисления

Операция Compute (Вычислить) позволяет путем вычислений создавать новые переменные на основе существующих данных файла. Так, среди данных файла ex01.sav исследователю может понадобиться средний балл результатов тестирования (тест1, ..., тест5) каждого учащегося. Для этого необходимо создать новую переменную, например, с именем тест_ср.

**Шаг 46** При активном окне редактора данных в меню Transform (Преобразование) выберите команду Compute (Вычислить). На экране появится диалоговое окно Compute Variable (Вычислить переменную), представленное на рис. 4.2.

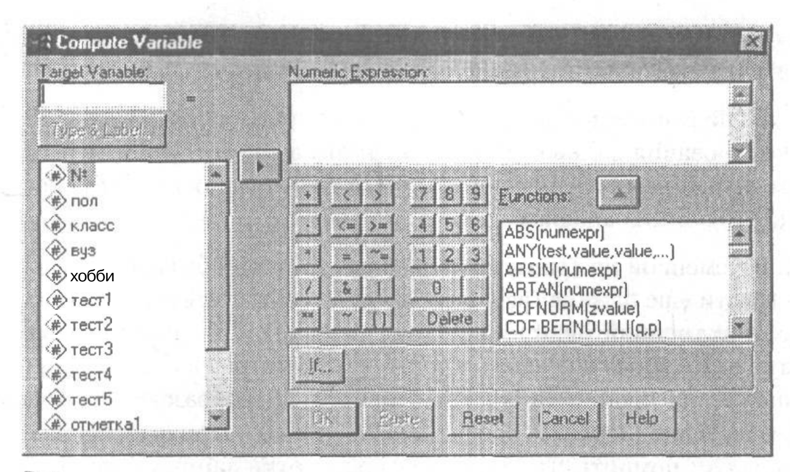



Рис. 4.2. Диалоговое окно Compute Variable

Диалоговое окно Compute Variable (Вычислить переменную) содержит в левой части уже ставший привычным список доступных переменных. Над списком, в левом верхнем углу окна, имеется поле Target Variable (Целевая переменная), в которое необходимо ввести имя создаваемой или вычисляемой переменной. В поле Numeric Expression (Выражение) вводится выражение, с помощью которого вычисляется новая переменная. Вы можете вводить в это поле символы с клавиатуры, через буфер обмена, а также пользоваться перечисленными ниже вспомогательными элементами интерфейса окна, позволяющими формировать выражение.

- На панели калькулятора содержатся кнопки цифр и десятичного разделителя, а также разнообразных операций. Щелчок на каждой из этих кнопок приводит к появлению соответствующего символа (или нескольких символов) в поле Numeric Expression (Выражение). Разумеется, при составлении выражения вы можете обойтись без панели калькулятора, поскольку все представленные там символы можно ввести с клавиатуры. Кнопки основных операций перечислены в табл. 4.1.
- В списке Functions (Функции) перечислены названия 70 функций, причем назначение большинства из них весьма туманно. Далее мы приводим описания десяти функций, которые кажутся нам наиболее употребительными. Аргумент *numexpr* означает численное выражение.
- При щелчке на кнопке If (Если) открывается диалоговое окно, почти идентичное окну Compute Variable (Вычислить переменную). Разница между ними состоит лишь в том, что первое дополнительно содержит пару переключателей Include all cases (Включить все объекты) и Include if case satisfies condition (Включить, если объект удовлетворяет условию), а также кнопку Continue (Продолжение).

Все вышеперечисленные элементы интерфейса будут использованы в последующем примере создания новой переменной тест_ср. Эта переменная получается путем сложения значений переменных тест1, ..., тест5 и делением суммы на 5.

**Таблица 4.1.** Кнопки основных операций, доступных при создании переменных

| Арифметические операции | Операции сравнения | Логические операции |
|---|---|---|
| +сложение | <меньше | &И |
| -вычитание | >больше | \|ИЛИ |
| *умножение | < = меньше или равно | ~ НЕ |
| /деление: | > = больше или равно | |
| "возведение в степень | = равно | |
| ()очередность операций | ~ = не равно | |

Далее перечислены основные функции, доступные при создании переменных.

ABS(*numexpr*)

Вычисляет модуль аргумента. Если zpositiv — целевая переменная, то выражение ABS(zscore) создаст переменную zpositiv, чьи значения будут равны модулям соответствующих значений переменной zscore.

RND(*numexpr*)

Возвращает значение аргумента, округленное до ближайшего целого. Если simple — целевая переменная, то выражение RND(отметка1) создаст переменную simple, чьи значения будут равны значениям переменной отметка1, округленным до ближайшего целого.

TRUNC(*numexpr*)

Возвращает целую часть аргумента. Если easy — целевая переменная, то выражение TRUNC(отметка1) создаст переменную easy, чьи значения будут равны значениям переменной отметка1 с отсеченной дробной частью.

MOD(*numexpr*, *modulus*)

Возвращает остаток от деления аргумента *numexpr* па число *modulus* (обычно равное 10). Если remain — целевая переменная, то выражение MOD(N, 10) создаст переменную remain, чьи значения будут равны последней цифре числа N. Если N = 86, то remain = 6.

SQRT(*numexpr*)

Возвращает квадратный корень аргумента. Если тест1rt — целевая переменная, то выражение SQRT(тест1) создаст переменную тест1rt, чьи значения будут равны корню квадратному от каждого значения тест1.

EXP(*numexpr*)

Возвращает число e (экспонента), возведенное в степень аргумента. Если confuse — целевая переменная, то выражение EXP(отметка1) создаст переменную confuse, чьи значения будут равны числу e в степени отметка1. $e \approx 2{,}721...$; $EXP(3{,}49) = (2{,}721...)^{3{,}49} \approx 32{,}900506$.

LG10(*numexpr*)

Возвращает десятичный логарифм аргумента. Если rhythm — целевая переменная, то выражение LG10(тест1) создаст переменную rhythm, чьи значения будут равны десятичным логарифмам переменной тест1.

LN(*numexpr*)

Возвращает натуральный логарифм аргумента. Если natural — целевая переменная, то выражение LG10(тест1) создаст переменную rhythm, чьи значения будут равны натуральным логарифмам переменной тест1.

NORMAL(s*tddev*)

Возвращает массив нормально распределенных случайных чисел с нулевым средним значением и стандартным отклонением, равным *stddev.* Если randnorm — целевая переменная, то выражение NORMAL(3) создаст переменную randnorm, значениями которой будут случайные числа, распределенные нормально с нулевым средним значением и стандартным отклонением, равным 3.

UNIFORM (*max*)

Возвращает массив случайных чисел, полученных в результате равновероятной выборки из целочисленного диапазона от 1 до *max.* Если random — целевая переменная, то выражение UNIFORM(100) создаст переменную random, чьи значения будут являться случайными целыми числами от 1 до 100.

Шаг *56* Для выполнения следующих действий диалоговое окно Compute Variable (Вычислить переменную) должно быть открыто (см. рис. 4.2).

1. В поле Target Variable (Целевая переменная) введите имя тест_ср.
2. Щелкните сначала на имени тест1, чтобы выделить его, а затем — на кнопке с направленной вправо стрелкой (можно также дважды щелкнуть на имени тест1), чтобы ввести его в поле Numeric Expression (Выражение).
3. Щелкните на кнопке + (или введите знак + с клавиатуры) и дважды щелкните па переменной тест2.
4. Повторите тс же действия для переменных тест3, тест4 и тест5.
5. Выделите правой кнопкой мыши все выражение, которое у вас получилось в результате выполнения шагов 2-4, и щелкните на кнопке с изображением открывающей и закрывающей скобок. В поле Numeric Expression (Выражение) сумма окажется в скобках.
6. Щелкните сначала на кнопке / (деление), затем на кнопке 5. В итоге должно получиться выражение: (тест1 + тест2 + тест3 + тест4 + тест5) / 5. Если это не так, внесите изменения. Если так, щелкните на кнопке OK.
7. В меню File (Файл) выберите команду Save (Сохранение).

В результате выполнения процедуры будет создана новая переменная тест_ср, которая разместится в крайнем правом столбце файла данных. Если вам удобнее держать данную переменную ближе к началу файла, можете воспользоваться операциями вырезания и вставки в нужную позицию.

После завершения процедуры программа возвращает вас в окно редактора данных, чтобы вы могли просмотреть содержимое новой переменной.

В примере использовались две арифметические операции: сложение и деление, при этом в составляемом выражении применялась очевидная очередность выполнения операций. В более сложных вычислениях с привлечением разнообразных арифметических и других операций вам придется следить за приоритетами операций и при необходимости регулировать очередность вычислений с помощью скобок.

## Ранжирование

Рассмотренная команда Compute (Вычислить) позволяет создать новую переменную, значения которой вычисляются при помощи выражения, содержащего другие нeременные. Команда Transform ▸ Rank Cases (Преобразование ▸ Ранжировать объекты) тоже позволяет создать новую переменную, значения которой — ранговые места объектов по заданной переменной. Эта процедура применяется, когда необходимо перейти от исходных значений переменной к рангам.

Команда ранжирования (перехода к рангам) выполняется при помощи диалогового окна, представленного на рис. 4.3. Для перехода к рангам достаточно задать имя переменной в этом окне и щелкнуть на кнопке ОК. Появится новая переменная с заданным по умолчанию именем, содержащая ранги значений исходной переменной.

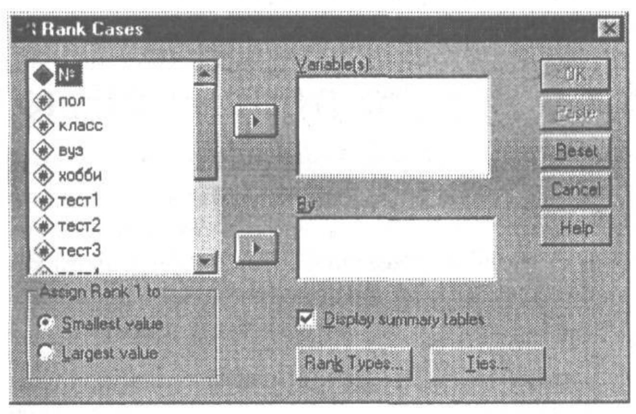



**Рис. 4.3.** Диалоговое окно Rank Cases

Предположим, исследователь решил ранжировать всех учащихся по успеваемости за 11-й класс. Для этого ему необходимо перейти от исходных значений переменной отметка2 к новой переменной, содержащей ранги учащихся по этой переменной.

Рассмотрим процедуру ранжирования на примере.

**Шаг 5в** Для ранжирования переменной отметка2 выполните следующие действия:

1. В меню Transform (Преобразование) выберите команду Rank Cases (Ранжировать объекты). На экране появится диалоговое окно, представленное на рис. 4.3.
2. Щелкните сначала на переменной отметка2, чтобы выделить ее, а затем — на кнопке со стрелкой.
3. Флажок Display summary tables (Показать таблицу вывода) можно сбросить, так как никаких результатов обработки, кроме новой переменной, получать не планируется.
4. По умолчанию в группе Assign Rank 1 to (Присвоить ранг 1) установлен переключатель Smallest value (Наименьшему значению). При желании вы можете изменить порядок ранжирования на обратный, установив в группе Assign Rank 1 to (Присвоить ранг 1) переключатель Largest value (Наибольшему значению).
5. При щелчке на кнопке Rank Types (Типы ранжирования) вам будет предложен дополнительный набор экстравагантных способов ранжирования, а при щелчке на кнопке Ties (Связи) — способов обработки одинаковых (связанных) рангов. По умолчанию принято обычное присвоение одинаковым значениям переменной одного и того же среднего ранга (Mean). Для ранжирования по общепринятым правилам оставьте все те параметры, которые установлены по умолчанию, и щелкните па кнопке OK.

В результате выполнения этого шага в конце списка появится новая переменная с именем rотметка, присвоенным по умолчанию. Это имя можно оставить или поменять па другое.

# Перекодировка в новую переменную

Команда Transform ▸ Recode ▸ Into Different Variables (Преобразование ▸ Перекодировка ▸ В другие переменные) создает новую переменную, однако ее значения определяются не как результат вычислений или ранжирования, а па основе замены множества значений существующей переменной небольшим числом категорий. Как правило, эта процедура применяется, когда необходимо разделить выборку объектов на подгруппы по некоторому количественному признаку.

Предположим, что, исходя из данных файла ex01.sav, требуется разделить учащихся на три группы по успеваемости за последний год (отметка2): 1 — слабо успевающие (примерно 30 %), 2 — средне успевающие (примерно 40 %) и 3 — отличники (примерно 30 %). Для этого можно использовать ранги, которые были получены путем ранжирования переменной отметка2 (новая переменная — rотметка).

Команда перекодирования в другие переменные выполняется с помощью двух диалоговых окон. Одно из них представлено па рис. 4.4. В нем вы можете задать входную и выходную переменные. Щелчок на кнопке Old and New Values

(Старые и новые величины) приведет к открытию диалогового окна, показанного на рис. 4.5. В нем вы можете задать градации новой переменной, которые будут соответствовать диапазонам уровней старой переменной.

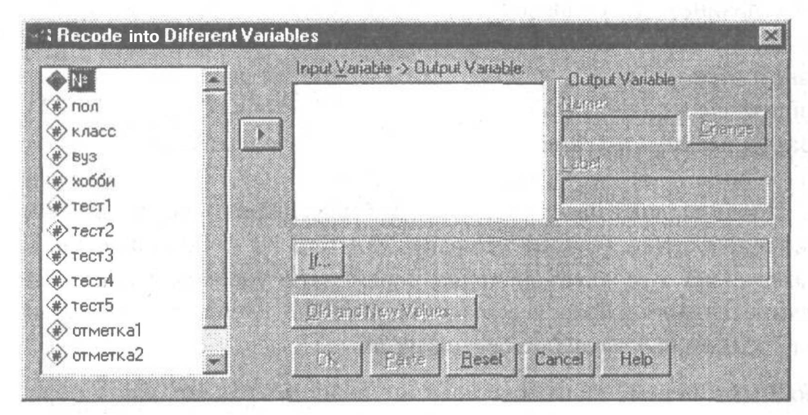



**Рис. 4.4.** Диалоговое окно Recode into Different Variables

Как видно на рис. 4.4, окно Recode into Different Variables (Перекодировка в другие переменные) имеет вполне обычный вид: в его левой части находится список доступных переменных, а внизу — пять стандартных кнопок (см. главу 2). Первое, что необходимо сделать в рассматриваемом нами примере, — выделить переменную готметка и переместить ее в список Input Variable -> Output Variable (Входная переменная -> Выходная переменная). Затем в поле Name (Имя) области Output Variable (Выходная переменная) следует ввести имя новой переменной (в данном случае уровень). Щелчок на кнопке Change (Изменить) приведет к появлению переменной уровень в предыдущем списке: его содержимое будет иметь вид готметка->уровень. Далее с помощью кнопки Old and New Values (Старые и новые величины) необходимо открыть диалоговое окно, показанное на рис. 4.5.

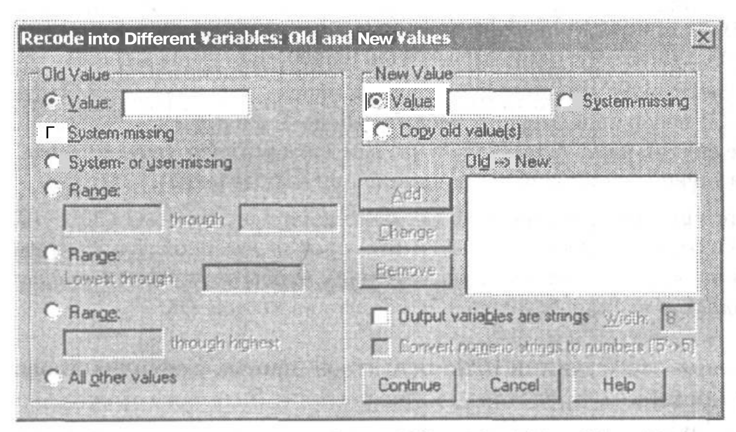



**Рис. 4.5.** Диалоговое окно Recode into Different Variables: Old and New Values

Теперь требуется задать соответствие между диапазонами значений переменной готметка и градациями повой переменной уровень. Для того чтобы настроить окно на работу с диапазонами, установите переключатель Range (Область) в группе Old Value (Старая величина). Далее введите в левое иоле, расположенное под переключателем Range (Область), число 1, в правое иоле — число 30, а в поле Value (Величина) в правом верхнем углу — цифру 1. После щелчка на кнопке Add (Добавление) вы увидите созданное соответствие в поле Old->New (Старая->Новая): 1 thru 30 -> 1. Затем введите в левое поле под переключателем Range (Область) число 30,1, в правое поле — число 70, а в поле Value (Величина) в нравом верхнем углу — цифру 2. После щелчка на кнопке Add (Добавление) вы увидите второе соответствие в поле Old->New (Старая->Новая): 30.1 thru 70 -> 2. Аналогично, вводя границы диапазона и нужные значения переменной уровень, задайте последнее соответствие. После того как все значения введены, щелкните сначала на кнопке Continue (Продолжить), затем — на кнопке OK.

Новыми производными переменными вы можете управлять так же, как и переменными, созданными «обычным» способом: назначать метки, вручную изменять значения, менять расположение в файле данных и т. п.

Обобщим процесс создания переменной уровень в форме уже привычных нам пошаговых процедур.

**Шаг 5г** Для создания переменной уровень выполните следующие действия:

1. В меню Transform (Преобразование) выберите команду Recode ▸ Into Different Variables (Перекодировка ▸ В другие переменные). На экране появится диалоговое окно, представленное па рис. 4.4.
2. Щелкните сначала на переменной готметка, чтобы выделить ее, а затем — на кнопке со стрелкой.
3. Для перехода в поле Name (Имя) нажмите клавишу Tab, введите имя уровень и щелкните на кнопке Change (Изменить).
4. Щелкните на кнопке Old and New Values (Старые и новые величины), чтобы открыть диалоговое окно, показанное па рис. 4.5.
5. В левой части окна установите четвертый сверху переключатель Range (Область). В левое поле под ним введите число 1, а в правое — число 30, щелкните внутри поля Value (Величина) в правом верхнем углу окна, введите туда цифру 1 и щелкните на кнопке Add (Добавление).
6. Повторите предыдущее действие для оставшихся диапазонов (30,1-70, 70,1–100) и соответствующих им значений новой переменной (2 и 3), затем щелкните на кнопке Continue (Продолжение), чтобы вернуться в предыдущее диалоговое окно, в котором щелкните па кнопке OK.

В результате выполнения этого шага в окне редактора данных появится новая переменная с именем уровень.

# Перекодировка существующей переменной

Иногда у вас может возникнуть необходимость изменить кодирование какой-либо переменной. Это может быть обусловлено двумя причинами.

- Вы работаете одновременно с несколькими файлами данных, которые создавались разными людьми и содержат одни и те же переменные, но закодированные по-разному.
- В процессе исследования вам стало ясно, что текущее кодирование какой-либо из переменных можно улучшить.

Обратимся к файлу ex01.sav. Как вы помните, в переменной пол значение 1 соответствует девушкам, а значение 2 — юношам. Вы можете столкнуться с тем, что у вас появится другой файл, в котором имеется переменная пол, но автор файла значением 1 закодировал юношей, а значением 2 — девушек. Чтобы избежать путаницы, вам придется перекодировать значения переменной пол в одном из двух файлов. Может оказаться и так, что значения, представляющие внешкольные увлечения учащихся, удобно расположить в порядке убывания распространенности этих увлечений среди школьников, и если вы не подумали об этом сразу, то вам также придется заняться перекодировкой.

Нередко при первоначальном кодировании вводится избыточное число уровней переменной. Как правило, эта избыточность проявляется на этапе проведения исследования. Типичной является следующая ситуация. Пусть имеется переменная вуз, отражающая выбираемый учащимся профиль образования. Первоначально задается 4 уровня этой переменной: 1 — гуманитарный, 2 — экономический, 3 – технический, 4 — естественнонаучный. При помощи процедуры, которая будет описана далее, можно объединить значения 1 и 2 в группу гуманитарных вузов, а значения 3 и 4 — в группу физико-технических вузов, и присвоить этим группам значения 1 и 2 соответственно.

Перекодировка существующей переменной выполняется с помощью команды Recode ▸ Into Same Variables (Перекодировка ▸ В те же переменные) меню Transform (Преобразование). После выбора этой команды на экране появится диалоговое окно, показанное на рис. 4.6.

Операция перекодирования в те же переменные, как и предыдущая, управляется двумя диалоговыми окнами (второе представлено на рис. 4.7), причем можно работать одновременно с несколькими переменными, однако для этого вам придется постоянно переключаться между диалоговыми окнами.

Первым действием, которое необходимо выполнить в окне Recode into Same Variables (Перекодировка в те же переменные), является заполнение списка перекодируемых переменных Variables (Переменные). Имена переменных содержатся в списке слева и выбираются при помощи кнопки со стрелкой. При щелчке на кнопке Old and New Values (Старые и новые величины) открывается диалоговое

окно, представленное на рис. 4.7. Далее описаны переключатели группы Old Value (Старая величина).

- Value (Значение). При установке этого переключателя в поле рядом нужно указать значение.
- System-missing (Физически пропущенное значение). Установка этого переключателя означает физическое отсутствие значения.

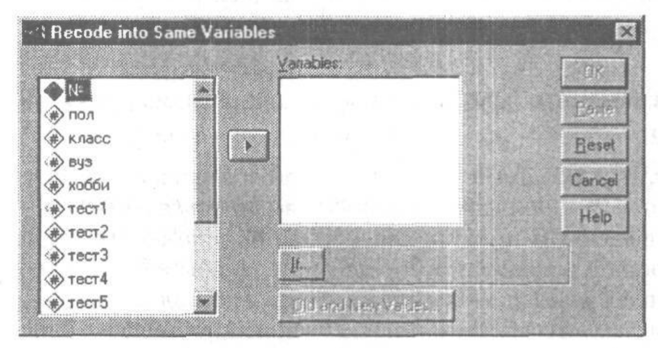



**Рис. 4.6.** Диалоговое окно Recode into Same Variables

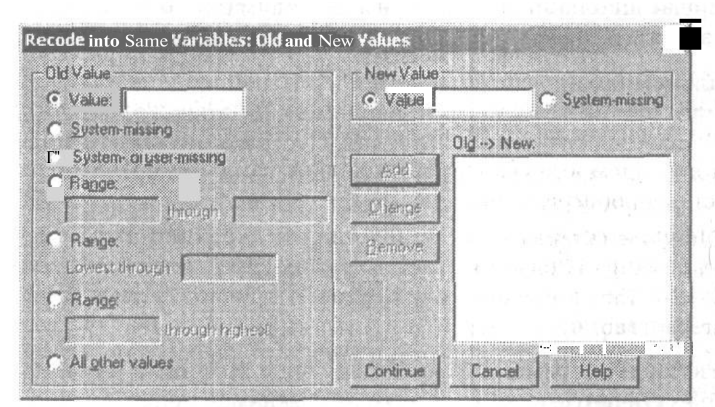



**Рис. 4.7.** Диалоговое окно Recode Into Same Variables: Old and New Values

- System- or user-missing (Физически или логически пропущенное значение). Установка этого переключателя означает, что значение либо не задано, либо не может участвовать в анализе.
- Range (Область). Это имя соответствует сразу трем переключателям, позволяющим тремя способами указать диапазон значений:
  - от нижней до верхней границы;
  - от наименьшего до заданного значения;
  - от заданного до наибольшего значения.

- All other values (Все другие значения). Этот переключатель устанавливается во всех остальных случаях.

Как и в предыдущей процедуре, вы задаете одно или несколько старых и новых значений и помещаете их в список Old->New (Старая->Новая). В приведенных ниже примерах мы поменяем местами два значения переменной пол и объединим 4 градации переменной вуз в две градации.

Шаг 5д Для перекодирования переменных пол и вуз выполните описанные ниже действия:

1. В меню Transform (Преобразование) выберите команду Recode > Into Same Variables (Перекодировка ▸ В те же переменные). На экране появится диалоговое окно, показанное на рис. 4.6.
2. Щелкните сначала на переменной пол, чтобы выделить ее, затем — на кнопке со стрелкой, чтобы переместить ее в список Variables (Переменные), и, наконец, на кнопке Old and New Values (Старые и новые величины), чтобы открыть диалоговое окно, показанное на рис. 4.7.
3. Введите в поле Value (Величина) области Old Value (Старая величина) число 1, дважды нажмите клавишу Tab, чтобы переключиться в поле Value (Величина) области New Value (Новая величина), введите в это поле число 2 и щелкните па кнопке Add (Добавление).
4. Повторите предыдущее действие для старого значения, равного 2, и нового значения, равного 1, а затем щелкните на кнопке Continue (Продолжение), чтобы вернуться в исходное диалоговое окно.
5. Теперь мы перейдем к перекодированию переменной вуз. Повторите второе действие для переменной вуз.
6. В группе Old Value (Старая величина) установите верхний из трех переключателей Range (Область), в левом поле введите значение 1, нажмите клавишу Tab, в правом поле введите значение 2, дважды нажмите клавишу Tab, чтобы перейти в поле Value (Величина) области New Value (Новая величина), введите там число 1 и щелкните на кнопке Add (Добавление).
7. Повторите предыдущее действие для старых значений 3-4 и нового значения 2, затем щелкните на кнопке Continue (Продолжение), чтобы вернуться в исходное диалоговое окно, в котором щелкните на кнопке OK.

Обратите внимание на то, что процедура перекодирования никак не влияет па метки значений. Другими словами, после выполнения перекодирования необходимо заново задать метки всем значениям, изменившим свой смысл. Действия с метками значений были рассмотрены в главе 3, и при необходимости вы можете обратиться к соответствующим разделам. Если вы хотите сохранить изменения, сделанные вами в файле данных, в меню File (Файл) выберите команду Save (Сохранение).

# Выбор объектов для анализа

В этом разделе мы отойдем от привычных предисловий и сразу перейдем к пошаговым процедурам.

**Шаг 4е** В меню Data (Данные) выберите команду Select Cases (Выбор объектов). На экране появится диалоговое окно, представленное на рис. 4.8.

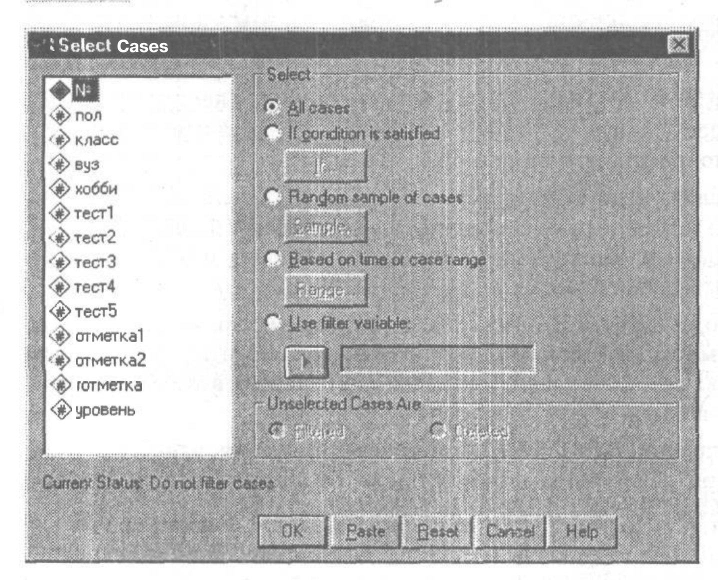
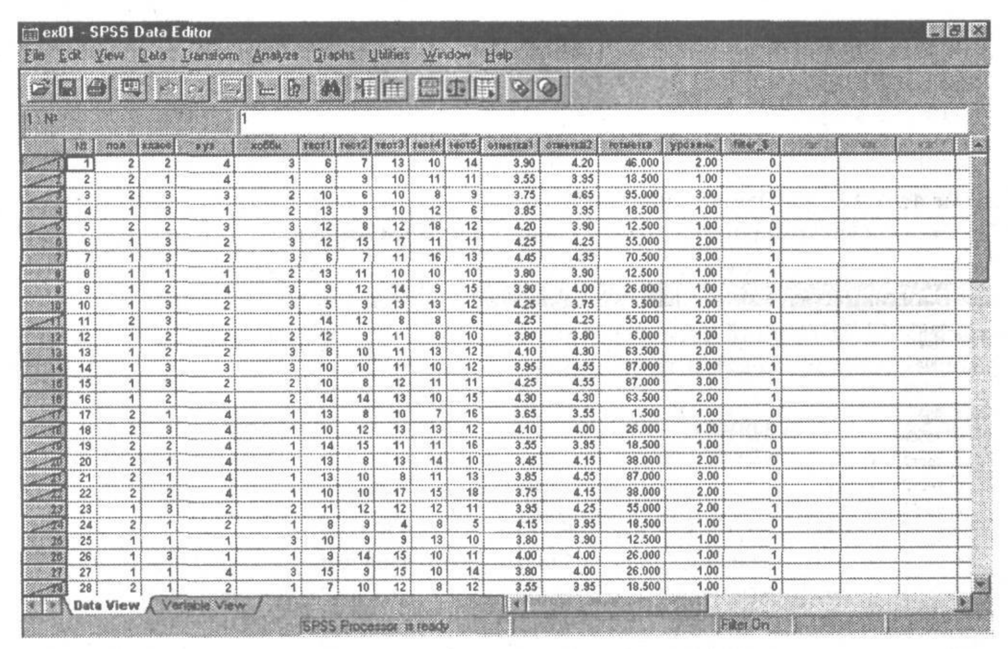



**Рис. 4.8.** Диалоговое окно Select Cases

Команда Select Cases (Выбор объектов) позволяет пользователю выбирать для обработки не все, а часть данных, удовлетворяющих заданным условиям. Поскольку необходимость в этом возникает довольно часто, команда Select Cases (Выбор объектов) является одной из самых востребованных при проведении исследований. Так, преподавателю или исследователю, использующему файл ex01.sav, могут понадобиться статистические сведения, касающиеся учащихся одного из классов или определенного пола, и т. д. Это означает, что ему нужно указать программе, какие данные следует выделить для обработки. Именно для этого предназначена команда Select Cases (Выбор объектов). После проведения анализа выбранной подгруппы вы можете вернуться к полному набору данных, установив в окне Select Cases (Выбор объектов) переключатель All Cases (Все объекты). Если вы хотите создать файл данных, включающий только выделенную их часть, можете удалить все исключенные объекты, а затем сохранить полученный файл с помощью команды Save (Сохранение) меню File (Файл). Номера выбранных объектов, в отличие от исключенных, после выполнения команды Select Cases (Выбор объектов) не зачеркиваются, что позволяет легко находить их визуально. На рис. 4.9 показана вкладка Data View (Просмотр данных) окна редактора данных, на которой видны выбранные объекты (девушки) и исключенные объекты (юноши).





| № | пол | класс | вуз | хобби | тест1 | тест2 | тест3 | тест4 | тест5 | отметка1 | отметка2 | [illegible] | уровень | filter_$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 2 | 4 | 3 | 6 | 7 | 13 | 10 | 14 | 3.90 | 4.20 | 46.000 | 2.00 | 0 |
| 2 | 2 | 1 | 4 | 1 | 8 | 9 | 10 | 11 | 11 | 3.55 | 3.95 | 18.500 | 1.00 | 0 |
| 3 | 2 | 3 | 3 | 2 | 10 | 6 | 10 | 8 | 9 | 3.75 | 4.65 | 95.000 | 3.00 | 0 |
| 4 | 1 | 3 | 1 | 2 | 13 | 9 | 10 | 12 | 6 | 3.85 | 3.95 | 18.500 | 1.00 | 1 |
| 5 | 2 | 2 | 3 | 3 | 12 | 8 | 12 | 18 | 12 | 4.20 | 3.90 | 12.500 | 1.00 | 0 |
| 6 | 1 | 3 | 2 | 3 | 12 | 15 | 17 | 11 | 11 | 4.25 | 4.25 | 55.000 | 2.00 | 1 |
| 7 | 1 | 3 | 2 | 3 | 6 | 7 | 11 | 16 | 13 | 4.45 | 4.35 | 70.500 | 3.00 | 1 |
| 8 | 1 | 1 | 1 | 2 | 13 | 11 | 10 | 10 | 10 | 3.80 | 3.90 | 12.500 | 1.00 | 1 |
| 9 | 1 | 2 | 4 | 3 | 9 | 12 | 14 | 9 | 15 | 3.90 | 4.00 | 26.000 | 1.00 | 1 |
| 10 | 1 | 3 | 2 | 3 | 5 | 9 | 13 | 13 | 12 | 4.25 | 3.75 | 3.500 | 1.00 | 1 |
| 11 | 2 | 3 | 2 | 2 | 14 | 12 | 8 | 8 | 6 | 4.25 | 4.25 | 55.000 | 2.00 | 0 |
| 12 | 1 | 2 | 2 | 2 | 12 | 9 | 11 | 8 | 10 | 3.80 | 3.80 | 6.000 | 1.00 | 1 |
| 13 | 1 | 2 | 2 | 3 | 8 | 10 | 11 | 13 | 12 | 4.10 | 4.30 | 63.500 | 2.00 | 1 |
| 14 | 1 | 3 | 3 | 3 | 10 | 10 | 11 | 10 | 12 | 3.95 | 4.55 | 87.000 | 3.00 | 1 |
| 15 | 1 | 3 | 2 | 2 | 10 | 8 | 12 | 11 | 11 | 4.25 | 4.55 | 87.000 | 3.00 | 1 |
| 16 | 1 | 2 | 4 | 2 | 14 | 14 | 13 | 10 | 15 | 4.30 | 4.30 | 63.500 | 2.00 | 1 |
| 17 | 2 | 1 | 4 | 1 | 13 | 8 | 10 | 7 | 16 | 3.65 | 3.55 | 1.500 | 1.00 | 0 |
| 18 | 2 | 3 | 4 | 1 | 10 | 12 | 13 | 13 | 12 | 4.10 | 4.00 | 26.000 | 1.00 | 0 |
| 19 | 2 | 2 | 4 | 1 | 14 | 15 | 11 | 11 | 16 | 3.55 | 3.95 | 18.500 | 1.00 | 0 |
| 20 | 2 | 1 | 4 | 1 | 13 | 8 | 13 | 14 | 10 | 3.45 | 4.15 | 38.000 | 2.00 | 0 |
| 21 | 2 | 1 | 4 | 1 | 13 | 10 | 8 | 11 | 13 | 3.85 | 4.55 | 87.000 | 3.00 | 0 |
| 22 | 2 | 2 | 4 | 1 | 10 | 10 | 17 | 15 | 18 | 3.75 | 4.15 | 38.000 | 2.00 | 0 |
| 23 | 1 | 3 | 2 | 2 | 11 | 12 | 12 | 12 | 11 | 3.95 | 4.25 | 55.000 | 2.00 | 1 |
| 24 | 2 | 1 | 2 | 1 | 8 | 9 | 4 | 8 | 5 | 4.15 | 3.95 | 18.500 | 1.00 | 0 |
| 25 | 1 | 1 | 1 | 3 | 10 | 9 | 9 | 13 | 10 | 3.80 | 3.90 | 12.500 | 1.00 | 1 |
| 26 | 1 | 3 | 1 | 1 | 9 | 14 | 15 | 10 | 11 | 4.00 | 4.00 | 26.000 | 1.00 | 1 |
| 27 | 1 | 1 | 4 | 3 | 15 | 9 | 15 | 10 | 14 | 3.80 | 4.00 | 26.000 | 1.00 | 1 |
| 28 | 2 | 1 | 2 | 1 | 7 | 10 | 12 | 8 | 12 | 3.55 | 3.95 | 18.500 | 1.00 | 0 |

**Рис. 4.9.** Просмотр данных после выбора объектов для анализа

При выборе группы объектов SPSS создает специальную переменную с именем filter_$, значение которой равно 1 для выбранных объектов и 0 для исключенных. Вы можете использовать эту переменную (например, для проведения сравнений между группами) или просто удалить ее.

В диалоговом окне Select Cases (Выбор объектов) имеются две группы переключателей. В группу Select (Выбор) входит 5 переключателей, из которых в данный момент нас интересует только два.

- Переключатель All cases (Все объекты) фактически предназначен для отмены операции, поскольку при его установке выбранными полагаются все объекты файла.
- Переключатель If condition is satisfied (Если удовлетворяет условию) позволяет задать условие отбора объектов с помощью кнопки If (Если). При щелчке на этой кнопке появляется диалоговое окно, очень похожее на окно, представленное на рис. 4.2. Чтобы задать условие отбора, вы можете использовать клавиатуру, панель калькулятора, список функций, а также буфер обмена. Далее приведены примеры некоторых условий отбора.

  ```
  пол = 1
  ```

  Отбираются только девушки.

  ```
  класс = 3
  ```

  Отбираются ученики класса «В».

  ```
  класс <= 3
  ```

Отбираются ученики классов «А» и «Б».

```
вуз >=2 & вуз <= 3
вуз > 1 & вуз < 4
```

При задании любого из этих условий отбираются ученики, выбирающие экономические и технические вузы.

Чтобы ввести переменную в поле задания условия отбора, достаточно щелкнуть сначала на ней, а затем — на кнопке со стрелкой. Знаки операций можно вводить как с клавиатуры, так и с панели калькулятора. В первом из примеров мы выберем для исследований всех учащихся-девушек, а во втором — учащихся, выбирающих экономические и технические вузы.

**Шаг 5е** Сначала займемся девушками. При открытом окне Select Cases (Выбор объектов), представленном на рис. 4.8, выполните следующие действия:

1. В группе Select (Выбор) установите переключатель If condition is satisfied (Если удовлетворяет условию) и щелкните на кнопке If (Если).
2. В открывшемся окне сначала щелкните па имени переменной пол, чтобы выделить ее, а затем на кнопке со стрелкой, чтобы ввести переменную пол в условие отбора. Далее на панели калькулятора щелкните на кнопке =, а потом — па кнопке 1, чтобы закончить составление условия отбора.
3. После создания условия отбора щелкните на кнопке Continue (Продолжение), чтобы закрыть первое диалоговое окно, и на кнопке OK, чтобы закрыть второе диалоговое окно и вернуться в окно редактора данных.

После выполнения этого шага при любой обработке будут учитываться только данные для девушек. Чтобы сделать доступными все данные, достаточно в окне Select Cases (Выбор объектов) установить переключатель All cases (Все объекты).

**Шаг 5е'** Теперь выберем для исследований школьников, ориентированных па поступление в экономические и технические вузы. После выполнения шага 5е окно Select Cases (Выбор объектов) должно выглядеть так, как показано на рис. 4.8.

1. В группе Select (Выбор) установите переключатель If condition is satisfied (Если удовлетворяет условию) и щелкните па кнопке If (Если).
2. В открывшемся окне сначала щелкните на имени переменной вуз, чтобы выделить ее, а затем — па кнопке со стрелкой, чтобы ввести переменную вуз в условие отбора. Далее на панели калькулятора щелкните па кнопке >=, введите число 2 и щелкните на кнопке &. Снова щелкните на имени переменной вуз, чтобы выделить ее, затем — на кнопке со стрелкой, чтобы во второй раз ввести переменную вуз в условие отбора. Теперь на панели калькулятора щелкните на кнопке <= и введите число 3, закончив составление условия отбора.
3. После создания условия отбора щелкните на кнопке Continue (Продолжение), чтобы закрыть первое диалоговое окно, и на кнопке OK, чтобы закрыть второе диалоговое окно и вернуться в окно редактора данных.

# Сортировка объектов

Команда Sort Cases (Сортировка объектов) предназначена для реорганизации данных файла. Эта операция очень распространена, поскольку позволяет расположить информацию в том порядке, в котором это удобно исследователю в текущий момент. Так, к примеру, исследователь мог бы с помощью этой команды изменить исходный порядок следования данных, чтобы учащиеся были перечислены но классам или по убыванию успеваемости (значению переменной отметка2). Подобные операции над данными можно выполнить с помощью команды Sort Cases (Сортировка объектов).

**Шаг 5ж** В меню Data (Данные) выберите команду Sort Cases (Сортировка объектов). На экране появится диалоговое окно, представленное на рис. 4.10.

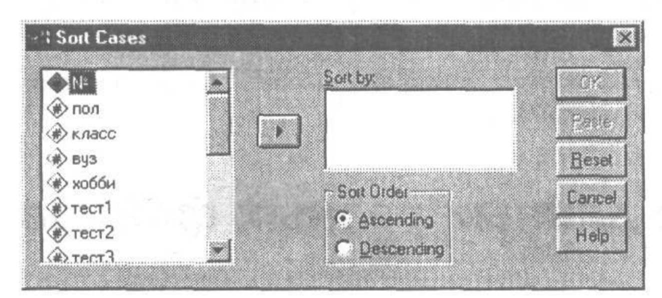



**Рис. 4.10.** Диалоговое окно Sort Cases

Параметры команды Sort Cases (Сортировка объектов), в отличие от параметров предыдущих команд, настраиваются в одном небольшом диалоговом окне. Вы выбираете переменные в исходном списке и с помощью переключателей Ascending (По возрастанию) или Descending (По убыванию) указываете порядок сортировки. Для строчных переменных (String) сортировка по возрастанию означает сортировку в алфавитном порядке.

При определении параметров сортировки можно задавать сразу несколько переменных. Это означает, что сначала данные будут отсортированы по значению первой выбранной переменной, затем объекты, имеющие одинаковые значения первой переменной, сортируются по значению второй переменной и т. д. Например, если в список Sort by (Сортировка) ввести переменные класс и пол, то все учащиеся будут перечислены по классам, а внутри каждого класса сначала будут следовать девушки, затем — юноши. Нередко результаты выполнения команды Sort Cases (Сортировка объектов) используются в качестве исходных данных для команды Cases Summaries (Сводка по объектам).

В первом из двух приведенных далее примеров данные файла ex01.sav упорядочиваются по убыванию значения переменной отметка2; во втором примере реализована упоминавшаяся ранее сортировка учащихся по классам и полу.

**Шаг 5з** При открытом диалоговом окне Sort Cases (Сортировка объектов), показан-
I ном на рис. 4.10, щелкните сначала на переменной отметка2, чтобы выделить ее, а затем — па кнопке со стрелкой, чтобы переместить ее в список Sort by (Сортировка). Далее в группе Sort Order (Порядок сортировки) установите переключатель Descending (По убыванию) и щелкните на кнопке OK.

**Шаг 5з'** При открытом диалоговом окне Sort Cases (Сортировка объектов), показанном на рис. 4.10, выполните следующие действия:

1. Щелкните сначала на переменной класс, чтобы выделить ее, а затем — па кнопке со стрелкой, чтобы переместить ее в список Sort by (Сортировка).
2. Повторите то же действие для переменной пол и щелкните на кнопке OK.

Обратите внимание на то, что по умолчанию выполняется сортировка по возрастанию, поэтому для такой сортировки никаких дополнительных действий по заданию способа сортировки не требуется.

Исходный порядок следования учащихся можно восстановить, назначив сортировку по переменной № (по возрастанию).

# Объединение данных разных файлов

Работа с данными нескольких файлов одновременно иногда способна сбить с толку даже опытных пользователей. Нередко файлы данных создаются при помощи разного программного обеспечения и имеют разные форматы. Это порождает различные проблемы их совместного использования, из которых мы рассмотрим наиболее типичную: дополнение *рабочего* файла SPSS содержимым *внешнего* файла. При этом возможны две ситуации дополнения данных рабочего файла SPSS:

- из внешнего файла Excel;
- из внешнего файла SPSS.

Мы рассмотрим оба варианта, но сначала сформулируем общие рекомендации:

- если вы намерены добавлять переменные, убедитесь, что порядок следования объектов в рабочем и внешнем файлах одинаков;
- если вы намерены добавлять объекты, убедитесь, что порядок следования переменных в рабочем и внешнем файлах одинаков;
- настройте форматы каждой переменной рабочего файла данных, чтобы они соответствовали данным внешнего файла (если добавляемая переменная содержит буквенные символы, то либо замените буквы числами во внешнем файле, либо поменяйте тип переменной па строчную в рабочем файле);
- перед добавлением данных создайте резервную копию рабочего файла, чтобы к ней можно было вернуться в случае неудачного переноса данных.

Следует иметь в виду, что чем больше соответствия между структурами рабочего и внешнего файлов, тем меньше вероятность ошибок при слиянии данных.

Если вы выполнили указанные рекомендации, то перенос данных из таблицы Excel не составит труда.

1. Откройте рабочий (SPSS) и внешний (Excel) файлы с данными.
2. В файле Excel выделите необходимый для переноса блок данных.
3. Скопируйте его в буфер обмена.
4. Разверните свернутое окно SPSS на вкладке Data View (Просмотр данных).
5. Установите фокус ввода в левую верхнюю ячейку того места, куда вы намереваетесь поместить переносимые данные.
6. В меню Edit (Правка) выберите команду Past (Вставка).

После этого таблица редактора данных дополнится новыми данными. Проверьте корректность переноса данных. Если допущена ошибка, в меню Edit (Правка) выберите команду Undo Modify Cell Values (Отмена изменения содержимого ячеек). После возврата в исходное состояние проверьте свои действия и повторите попытку.

При помощи той же последовательности действий можно перенести данные в рабочий файл из внешнего файла SPSS. Отличие заключается в том, что одновременно и рабочий, и внешний файлы не могут быть открытыми. Поэтому сначала следует открыть внешний файл, скопировать данные, затем найти и открыть рабочий файл, в который вставить скопированные данные.

Указанная последовательность действий требует большого внимания и определенной сноровки. Поэтому вряд ли этот метод может быть рекомендован не очень искушенным пользователям. К счастью, существует более надежный способ объединения данных — *слияние* рабочего и внешнего файлов.

Слияние файлов допустимо, когда и рабочий, и внешний файлы созданы при помощи редактора данных SPSS и имеют одинаковые имена переменных (когда мы добавляем объекты) и одинаковые число и порядок следования объектов (когда мы добавляем переменные). Таким образом, если внешним является файл Excel, то перед слиянием необходимо на его основе создать внешний файл SPSS (см. главу 3). Структура создаваемого файла должна быть максимально согласована со структурой рабочего файла. Для этого, дополнительно к общим рекомендациям, изложенным ранее, нужно настроить форматы каждой переменной файлов так, чтобы они были одинаковыми, и проверить идентичность имен переменных файлов.

Хотя эти рекомендации обязательными не являются, все же следует иметь в виду, что чем больше соответствия между структурами файлов, тем меньше вероятность ошибок при слиянии. Особое внимание следует уделять идентичности имен переменных, их форматам и расположению в файле данных. Обратите внимание также на то, что при добавлении переменных одинаковый порядок следования объектов в файлах обязателен.

## Добавление объектов

После завершения шага 3, описанного ранее в этой главе, у вас уже должен быть открыт файл ex01 .sav.

Шаг 5и Для добавления объектов в меню Data (Данные) выберите команду Merge Files ▸ Add Cases (Слияние файлов ▸ Добавить объекты). На экране появится диалоговое окно, представленное па рис. 4.11.

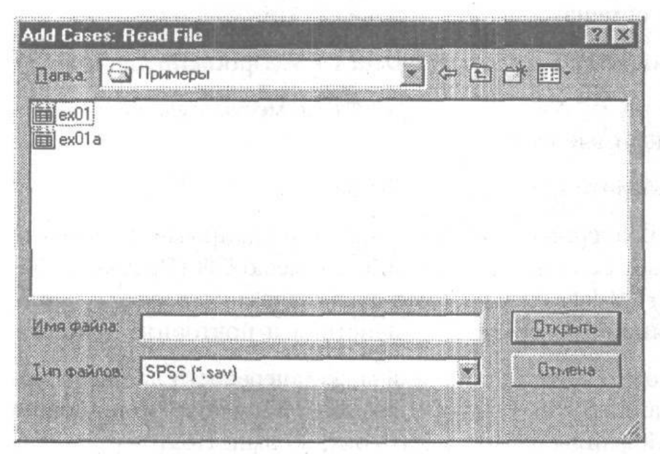



**Рис. 4.11.** Диалоговое окно Add Cases: Read File

С помощью диалогового окна Add Cases: Read File (Добавить объекты: чтение файла) можно выбрать внешний файл данных, предназначенный для слияния с открытым (рабочим) файлом. Если нужный внешний файл есть в списке диалогового окна, то для доступа к нему достаточно щелкнуть на его имени.

Предположим, создан файл ex01a.sav, в котором содержатся данные о 10 новых учащихся с теми же переменными, что и файл ex01.sav. В данном случае ex01.sav является рабочим файлом, а ex01a.sav — внешним. Иногда внешний файл располагается па гибком диске. Как правило, в этом случае проще ввести полный путь к файлу с клавиатуры в формате a:*имя_файла*.sav. После задания имени файла щелкните на кнопке Open (Открыть).

После того как внешний файл задан, па экране появляется диалоговое окно, показанное па рис. 4.12. Как видите, в заголовке окна отображено полное имя внешнего файла.

Список Unpaired Variables (Непарные переменные) в данном случае пуст (как и следовало ожидать). Если же файлы содержат непарные переменные, то все они попадают в список, при этом переменные рабочего файла помечаются знаком * (звездочка), а переменные внешнего файла — знаком + (плюс). После того как вы убедитесь, что структуры рабочего и внешнего файлов совпадают, необходимо проверить на идентичность структуру каждой пары переменных. Перед

щелчком на кнопке OK вы можете при необходимости выполнить два дополнительных действия.

- Если вы не хотите, чтобы какая-либо из парных переменных присутствовала в новом файле, выделите ее щелчком в списке Variables in New Working Data File (Переменные в новом рабочем файле) и удалите, нажав клавишу Del.

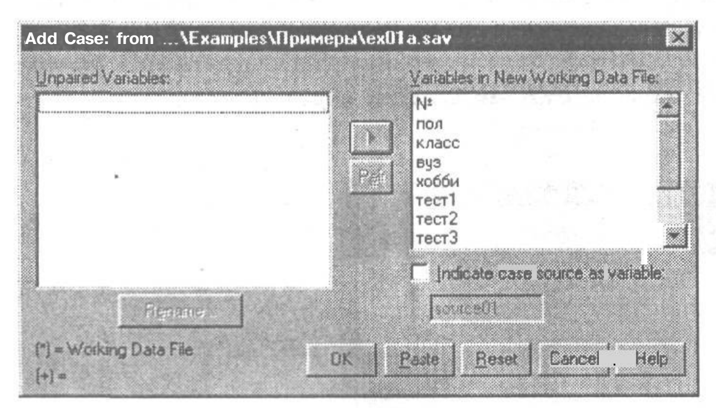



**Рис.** 4.12. Диалоговое окно Add Cases From File

- По умолчанию все непарные переменные исключаются из нового рабочего файла. Если же вы хотите включить какую-либо из непарных переменных в рабочий файл, то выделите ее в списке Unpaired Variables (Непарные переменные) и переместите в список Variables in New Working Data File (Переменные в новом рабочем файле) щелчком па кнопке со стрелкой. Переменная войдет в рабочий файл, а ее отсутствующие значения будут восприниматься как пропущенные.

**Шаг 5к** Чтобы выполнить слияние файлов ex01.sav и ex01a.sav, в диалоговом окне, представленном на рис. 4.11, щелкните сначала на имени ex01a.sav, потом — па кнопке Open (Открыть), и в открывшемся диалоговом окне (см. рис. 4.12) — на кнопке OK.

После выполнения этого шага в редакторе данных оказывается открытым новый рабочий файл с добавленными объектами, которому необходимо дать имя и сохранить. Содержимое старого рабочего файла и внешнего файла остается прежним.

# Добавление переменных

После завершения предыдущей пошаговой процедуры вам необходимо заново открыть файл ex01.sav.

**Шаг 5л** В меню Data (Данные) окна редактора данных выберите команду Merge Files ▸ Add Variables (Слияние файлов ▸ Добавить переменные). На экране появится диалоговое окно, представленное ранее па рис. 4.11, только заголовок окна будет другим.

Слияние с добавлением переменных очень похоже на слияние с добавлением объектов. Начало выполнения обеих пошаговых процедур фактически одинаково: па экране появляется диалоговое окно, в котором вы указываете имя внешнего файла и щелкаете на кнопке Open (Открыть). При добавлении переменных после этого открывается диалоговое окно, представленное на рис. 4.13. В его заголовке отражено полное имя внешнего файла. Обратите внимание, что для демонстрации дайной процедуры мы создали еще один специальный файл ex01b.sav, который содержит переменные идентификационного номера и класса для тех же объектов (учащихся), что и в файле ex01.sav, а также новую переменную тест6 – результат дополнительного тестирования.

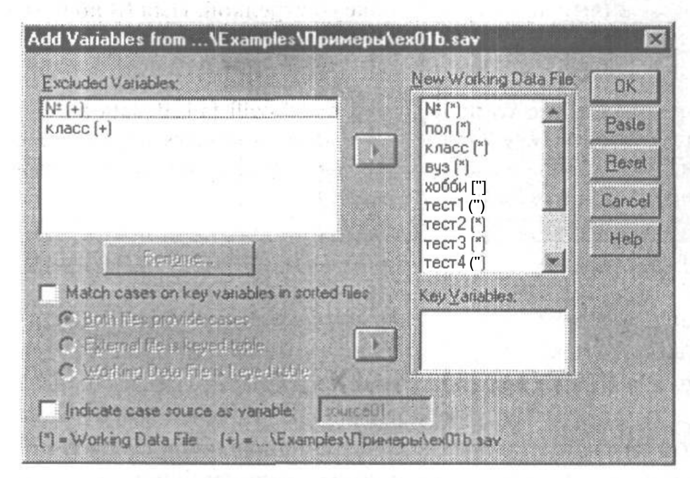



**Рис. 4.13.** Диалоговое окно Add Variables from File

Парные переменные попадают в список Excluded Variables (Исключенные переменные) и помечаются символом + (плюс), указывающим па то, что они принадлежат внешнему файлу. Переменные, включенные в новый рабочий файл, перечисляются в списке New Working Data File (Новый рабочий файл данных) и помечаются символом * (звездочка).

Следующим и очень важным действием является задание ключевой переменной. Необходимо, чтобы порядок следования значений в ключевых переменных обоих файлов был одинаков. В нашем примере в роли ключевой переменной рекомендуется использовать переменную № (идентификационный номер), поскольку ее значения уникальны для каждого объекта. Значения переменной № должны быть упорядочены, например по возрастанию. После того как вы убедитесь, что в обоих файлах ключевые переменные отсортированы одинаково, установите флажок Match cases on key variables in sorted files (Соответствие объектов по ключевым переменным в сортируемых файлах), выделите ключевую переменную в списке Excluded Variables (Исключенные переменные) и щелкните сначала на нижней кнопке со стрелкой, а йотом — па кнопке OK.

Ниже приведен пример слияния рабочего файла ex01.sav с внешним файлом ex01b.sav, где в качестве ключевой переменной выступает переменная №, упорядоченная по возрастанию.

**Шаг 5м** В диалоговом окне Add Variables: Read File (Добавить переменные: чтение файла), открытом после выполнения шага 5и, выполните следующие действия:

1. Щелкните сначала на имени ex01b.sav, а потом — на кнопке Open (Открыть). На экране появится диалоговое окно, показанное на рис. 4.13.
2. Установите флажок Match cases on key variables in sorted files (Соответствие объектов по ключевым переменным в сортируемых файлах), в списке Excluded Variables (Исключенные переменные) щелкните на переменной №, чтобы выделить ее, и щелкните на нижней кнопке со стрелкой. Имя № появится в списке Key Variables (Ключевые переменные).
3. Щелкните на кнопке OK.
4. Появится предупреждение Warning: Keyed match will fail, if data are not sorted in ascending order of Key Variables (Внимание: ключевое соответствие будет нарушено, если данные не отсортированы в порядке возрастания ключевой переменной). Щелкните на кнопке OK.

Если у вас возникнут проблемы при выполнении этого примера, проверьте, действительно ли переменная № упорядочена по возрастанию в обоих файлах.

# Печать результатов и выход из программы

Ниже описана типичная процедура печати результатов обработки данных (или серии обработок). Как вы знаете, результаты анализа помещаются программой в специальное окно вывода, поэтому при выполнении указанных инструкций необходимо предварительно переключиться в него, если это требуется.

**Шаг 6** В окне вывода укажите фрагменты, выводимые на печать (см. раздел «Окно вывода» в главе 2), в меню File (Файл) выберите команду Print (Печать), при необходимости задайте параметры печати и щелкните на кнопке OK.

Последнее, что необходимо сделать после завершения исследования и печати результатов, — это выйти из программы SPSS.

**Шаг 7** Для выхода из программы в меню File (Файл) выберите команду Exit (Выход).

Иногда после выполнения команды Exit (Выход) на экране могут появляться небольшие диалоговые окна с вопросом о необходимости сохранения сделанных в файлах изменений и кнопками, описывающими возможные варианты ответа. Для завершения работы просто щелкайте на соответствующих кнопках.

# 5 Диаграммы



Пакет SPSS for Windows обладает обширным арсеналом мощных и эффективных средств построения диаграмм. Разумеется, в книгу такого объема, как эта, бессмысленно даже пытаться вместить описания всех этих средств. Мы лишь коснемся основных аспектов работы с графическими объектами и при этом рассмотрим:

- команды построения диаграмм SPSS;
- команды редактирования диаграмм;
- пример диаграммы, с помощью которого вы познакомитесь с основными терминами, применяемыми в SPSS при описании графических объектов;
- диалоговое окно, характерное для большинства команд построения диаграмм в SPSS[1].

## Графика в программе SPSS

Как правило, диаграмма зависит от конкретной статистической процедуры, следовательно, рассматривать все тонкости построения диаграмм имеет смысл лишь в контексте обработки данных. Поэтому вся специфическая информация относительно конкретных диаграмм излагается по мере необходимости в главах книги, посвященных статистическому анализу. А здесь собраны лишь общие сведения, касающиеся диаграмм вообще, и по большей части эти сведения связаны с редактированием уже созданных диаграмм.

---

[1] Установленный по умолчанию шрифт для диаграмм SPSS, как правило, некорректно отображает русскоязычные наименования переменных и их меток. Для решения проблемы в меню Edit (Редактирование) выберите команду Options (Параметры), в открывшемся диалоговом окне перейдите на вкладку Charts (Диаграммы) и в списке Font (Шрифт) выберите тот шрифт, который допускает отображение кириллицы. Обычно это шрифты с окончанием Cyr, например Arial Cyr, Times New Roman Cyr и др.

Далее перечислены далеко не все, но наиболее часто используемые виды графиков.

- *Столбиковые диаграммы* (bar graphs), или гистограммы для дискретных переменных. Эти графики применяются для отображения распределения частот встречаемости значений переменной, имеющей небольшое количество градаций. Например, с помощью столбиковой диаграммы удобно представить распределение учащихся по трем классам или по их ориентации на поступление в вузы четырех типов, и т. д. Столбиковые диаграммы делятся на простые, кластерные и стековые.
- *Гистограммы* (histograms) внешне напоминают столбиковые диаграммы, однако, как правило, иллюстрируют распределение объектов по диапазонам значений непрерывной переменной (имеющей большое число возможных значений). С помощью гистограммы было бы удобно представить распределение учащихся по диапазонам значений успеваемости (отметки) или диапазонам тестовых значений.
- *Линейные графики* (line graphs) применяются для иллюстрации различных зависимостей между данными.
- *Круговые диаграммы* (pie charts), как и столбиковые, зачастую применяются для иллюстрации распределений в различных категориях.
- *Парето-диаграммы* (pareto charts) сочетают свойства столбиковых диаграмм и линейных графиков; при этом столбики представляют число элементов данных в различных категориях, а линии показывают накопленные частоты.
- *Коробчатые диаграммы* (box plots) основаны на процентилях и являются прекрасным средством отображения распределения данных.
- *Диаграммы рассеивания* (scatter plots) часто используются для отображения корреляций между переменными. Различают простые и оверлейные диаграммы рассеивания.
- *Диаграммы столбцов ошибок* (error bar charts) включают столбцы, отображающие стандартную погрешность измерения или доверительный интервал для сравниваемых групп.

SPSS предоставляет дополнительную очень удобную возможность построения графиков в диалоговом режиме. В меню Graphs (Графики) имеется подменю Interactive (Интерактивные), в котором представлен широкий спектр команд, позволяющих строить диаграммы в интерактивном режиме: добавлять переменные, изменять категории данных и т. п. Другими словами, работая исключительно с графиками, вы можете добиться такого же результата, как если бы применили статистические процедуры. Иногда такая возможность может оказаться полезной для пользователя, однако до этого необходимо освоить стандартные приемы работы с SPSS как в отношении статистики, так и в отношении графики. Поэтому описание процедур интерактивного построения диаграмм выходит за рамки

темы этой книги. Предполагается, что пользователь, набравший некоторый опыт работы с программой SPSS, сможет в дальнейшем с легкостью экспериментировать с интерактивными графиками.

# Пример диаграммы

На рис. 5.1 приведена диаграмма, которую мы постарались создать так, чтобы на ее примере проиллюстрировать максимально возможное число понятий, необходимых при работе с графическим средствами SPSS. Конечно, нам не удалось охватить одним примером все нужные термины, однако если в процессе работы вам понадобятся пояснения, касающиеся того или иного понятия, большую часть ответов на ваши вопросы вы сможете найти здесь. Обратите внимание, что вводимая терминология не является универсальной; напротив, она носит скорее локальный характер и предназначена для использования в рамках данной программы. Например, понятие *ось значений* носит для нас вполне определенный смысл, который может не совпадать со смыслом, вкладываемым в это понятие в других статистических или графических программах.

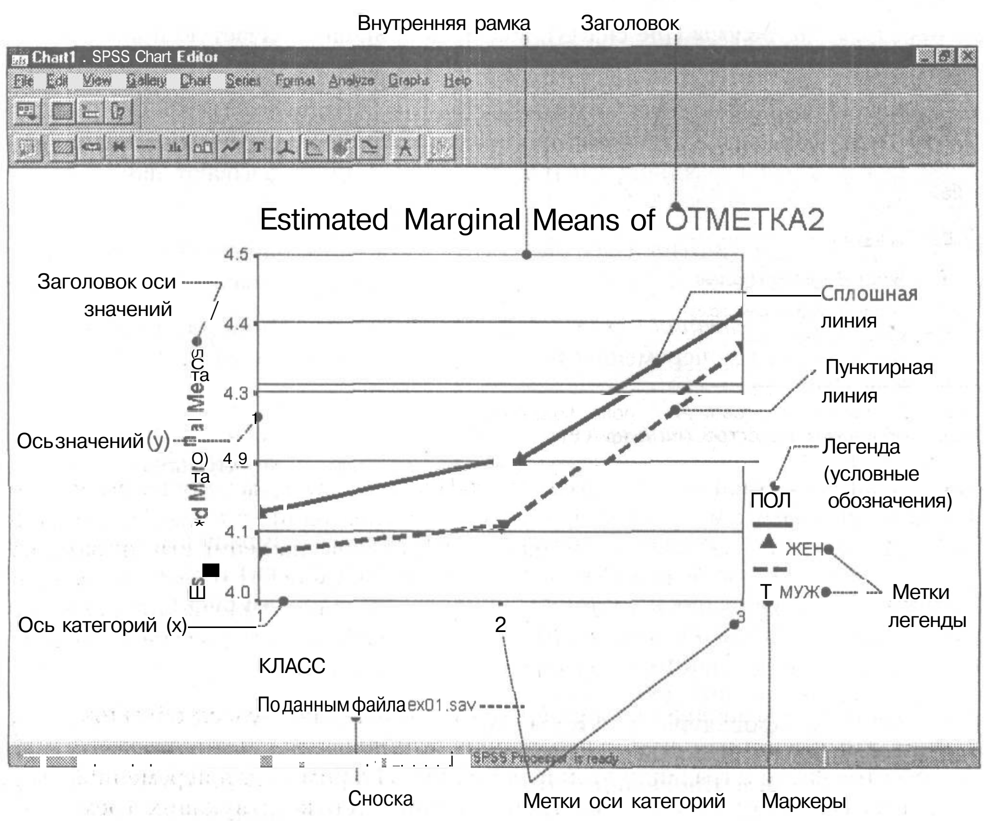



**Рис. 5.1.** Пример диаграммы

# Диалоговые окна команд построения графиков

Большинство статистических процедур, описанных в последующих главах книги, имеют встроенные средства построения графиков. Тем не менее программа SPSS предоставляет возможность разделения задач построения графиков и задач статистического анализа. Вы можете с равным успехом отредактировать готовый график, полученный в результате применения статистической процедуры, или новый график, созданный при помощи меню Graphs (Графики). В этой главе нет пошаговых процедур, однако рекомендации, которые вы здесь найдете, позволят вам эффективно работать с диаграммами. Мы опишем два основных типа диалоговых окон, которые используются графическими процедурами SPSS.

Диалоговое окно первого типа можно увидеть после выбора одной из команд построения диаграмм. На рис. 5.2 представлено диалоговое окно, появляющееся на экране при выборе команды Bar (Столбец) в меню Graphs (Графики).

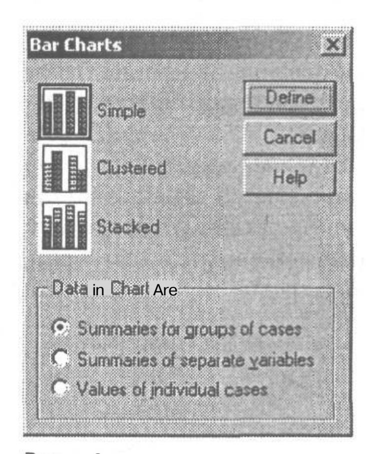



**Рис. 5.2.** Типичное диалоговое окно, появляющееся после выбора команды построения графика

Как правило, с помощью элементов интерфейса верхней части окна вы можете выбрать тип диаграммы — в данном случае доступен один из типов Simple (Простой), Clustered (Кластерный) и Stacked (Стековый), а в нижней части указать, какие данные вы намерены использовать. Типы диаграмм будут меняться в зависимости от вида диалогового окна; для того чтобы указать нужный тип, щелкните на кнопке с миниатюрой диаграммы слева от ее названия. Варианты используемых данных для большинства типов диаграмм одни и те же.

- Переключатель Summaries for groups of cases (Сводка по группам объектов) означает, что задействуется одна переменная, а столбцы диаграммы отразят число объектов каждой градации этой переменной. Например, для переменной класс диаграмма будет состоять из трех столбцов, соответствующих трем классам учащихся; для переменной пол — из двух столбцов, отображающих число

учащихся женского и мужского иола; для переменной вуз — из четырех столбцов, представляющих количество учащихся в каждой из четырех групп предпочтительных вузов.

- Переключатель Summaries of separate variables (Сводка по отдельным переменным) указывает на то, что диаграмма будет содержать несколько столбцов, каждый из которых будет соответствовать среднему значению одной из переменных. Как правило, при построении диаграмм этого типа используют переменные, связанные между собой по смыслу. В файле ex01.sav такими переменными являются переменные тест1. . . .тест5: диаграмма, построенная для них, отразит средние баллы всех учащихся для каждого из пяти тестов.
- Переключатель Values of individual cases (Значения для каждого объекта) предназначен для файлов данных с относительно небольшим числом объектов. Если взять файл ex01.sav и построить подобную диаграмму для какой-либо переменной, например тест1, то па экране появились бы 100 столбцов, отображающих баллы каждого из учащихся по первому тесту. Разумеется, если число объектов файла велико, то можно выбрать для обработки их подмножество (см. главу 4).

На рис. 5.3 показана характерная часть «вторичных» диалоговых окон большинства команд построения диаграмм, появляющихся при щелчке па кнопке Define (Определить). Как вы видите, это группа из пяти переключателей. Чаще других используются переключатели N of cases (Количество объектов) и % of cases (% объектов) — каждый столбец или сегмент графика представляет число (процент) объектов в каждой из категорий. Несколько реже используются переключатели Cum. n of cases (Накопленное количество объектов) и Cum. % of cases (Накопленный % объектов) — первый столбец представляет число объектов в первой категории, второй — число объектов в первой и второй категориях, третий — в первой, второй и третьей категориях и т. д.

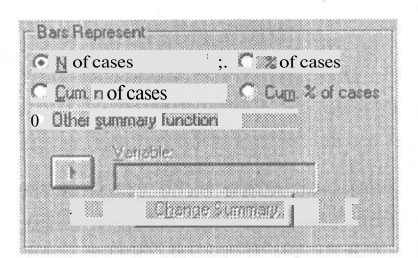



**Рис. 5.3.** Переключатели выбора варианта представления данных для диаграммы

На рис. 5.4 показаны две диаграммы, построенные для переменной хобби: диаграмма слева строилась при установленном переключателе N of cases (Количество объектов), а диаграмма справа — при установленном переключателе Cum. n of cases (Накопленное количество объектов). Переключатели % of cases (% объектов) и Cum. % of cases (Накопленный % объектов) аналогичны описанным выше с той лишь разницей, что по вертикальной оси диаграммы откладывается не количество объектов в категории, а их процент от общего числа объектов.

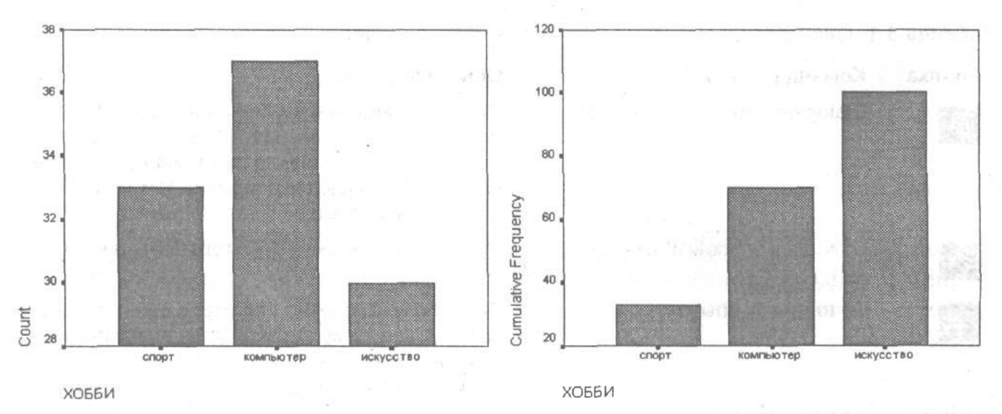



**Рис. 5.4.** Диаграммы, построенные с использованием количества объектов (слева) и накопленного количества объектов (справа)

# Редактирование графиков и диаграмм

После того как диаграмма создана, программа предоставляет вам широкий набор команд ее редактирования. Перед тем как выполнять какое-либо редактирование графического объекта, необходимо дважды щелкнуть на нем мышью. При этом па экране появится окно графического редактора, содержащее строку меню с полным набором команд и панель инструментов, на которой находятся кнопки, соответствующие наиболее часто используемым командам. Описания кнопок панели инструментов приведены в табл. 5.1.

## Процедура редактирования

Команды редактирования могут применяться к диаграмме только в том случае, если она отображена на экране в режиме редактирования. Как только вы дважды щелкнете на диаграмме, SPSS откроет новое окно с ее изображением и строкой меню в верхней части; это означает, что диаграмма доступна для редактирования. Строка меню (см. ниже) расположена над панелью инструментов и содержит полный перечень команд редактирования диаграмм.




Как уже упоминалось, для редактирования диаграмм используются либо команды меню, либо кнопки панели инструментов. Возможны две ситуации. В первом случае выбор команды или щелчок на кнопке сразу приводит к изменению диаграммы. Примером могут служить команда и кнопка Swap axes (Смена осей). Во втором случае на экране появляется диалоговое окно, предназначенное для задания дополнительных параметров команды. Как правило, параметры могут применяться к диаграмме одним из трех способов.

Таблица 5.1. Кнопки панели инструментов графического редактора

| Кнопка | Команда | Описание |
|---|---|---|
| | Dialog recall (Повтор диалога) | Позволяет вернуться к первоначальным установкам или вычислениям, не выходя из окна диаграммы. Как правило, эта команда применяется для возвращения к параметрам текущей диаграммы с целью ее редактирования |
| | Go to data window (К данным) | Переключение в окно редактора данных SPSS |
| | Go to case (К объекту) | Переключение в окно редактора данных с переходом к текущему объекту (если диаграмма не отображает величины для отдельных объектов, то осуществляется переход к первому объекту) |
| | Variables (Переменные) | Позволяет: получить доступ к параметрам переменных |
| | Point ID (Номер объекта для точки) | При щелчке на точке диаграммы распределения определяет соответствующий ей объект (если диаграмма отражает величины для отдельных объектов) |
| | Fill patterns (Образец заливки) | Служит для выбора готовых образцов заливки для элементов диаграммы (маркера, столбца) |
| | Color palette (Палитра цветов) | Изменение цвета маркера и цвета заливки столбца или сегмента |
| | Marker style (Стиль маркера) | Изменение формы и размера маркера графика |
| | Line style (Стиль линии) | Задание толщины и типа линии |
| | Bar style (Стиль столбца) | Создание эффекта тени или объема |
| | Bar labels (Метки столбцов) | Создание метки, отображающей значение для столбца |
| | Interpolation line (Линия интерполяции) | Задание стиля соединения точек графика и изображения маркера |
| | Font type and size (Тип и размер шрифта) | Определение шрифта текста. Набор шрифтов зависит от возможностей вашего компьютера |
| | 3D scatterplot rotation (Параметры 3D-вращения) | Задание параметров трехмерного вращения |
| | Swap axes (Смена осей) | Обмен местами вертикальной и горизонтальной осей |
| | Explode pie slice (Удаление сегмента) | Удаление одного или нескольких сегментов круговой диаграммы |
| | Break line at missing (Разрыв графика в утерянных точках) | Разрыв линии в графике в случае отсутствия категории |
| | Chart options (Параметры диаграммы) | Задание дополнительных параметров, специфичных для данного типа диаграммы |
| | 3D scatterplot spin mode (Режим 3D-вращения) | Включение режима вращения трехмерной диаграммы |

- При установке параметра сразу происходит изменение диаграммы, после чего достаточно щелкнуть па кнопке Close (Закрыть), чтобы закрыть диалоговое окно редактирования.
- Чтобы установленные параметры вступили в силу, требуется щелкнуть на кнопке Apply (Применить) или Apply to all (Применить ко всем). Происходит изменение диаграммы, после чего вы можете закрыть диалоговое окно редактирования щелчком на кнопке Close (Закрыть).
- Кнопки Apply (Применить) и Apply to all (Применить ко всем) недоступны. Это может быть обусловлено двумя причинами: либо параметр к данной диаграмме действительно неприменим, либо для выполнения операции необходимо предварительно выделить какой-либо фрагмент диаграммы. Последняя ситуация весьма распространена: выделение требуется для изменения шрифта текста, для изменения стиля маркера или линии, для удаления сегмента круговой диаграммы и т. д. Программа SPSS обеспечивает весьма удобный механизм работы с диаграммами: вы можете выполнять выделение как до выбора команды построения диаграммы, так и после него; при этом если вы сначала выбрали команду, то в процессе выделения фрагмента диаграммы диалоговое окно соответствующей команды остается открытым.

Некоторые команды построения диаграмм можно выполнять и до появления диаграммы на экране (например, задание меток и заголовков), однако это не является обязательным, и вы можете вернуться к подобным процедурам в любой момент. В конце главы мы приводим краткое описание наиболее важных команд подобного рода. Большинство из них можно выполнить с помощью кнопок панели инструментов, но иногда для их вызова проще использовать строку меню.

## Команды редактирования

Фактически мы рассмотрим команды лишь двух меню — Gallery (Галерея) и Chart (Диаграмма). Команды меню File (Файл), Edit (Редактирование), Window (Окно) и Help (Справка) уже знакомы вам по предыдущим главам, поэтому мы не станем сопровождать их дополнительными пояснениями.

Команды меню Gallery (Галерея) предназначены смены стилей диаграмм. Очевидно, что данные, представленные с помощью столбиковой диаграммы, могут быть также представлены с помощью линейного графика и других видов диаграмм. Было бы весьма удобно иметь возможность сначала построить диаграмму, а затем выбрать наиболее удобную форму отображения данных. SPSS предоставляет такую возможность: достаточно выбрать в меню Gallery (Галерея) команду задания типа диаграммы, затем в открывшемся диалоговом окне задать ее стиль и щелкнуть на кнопке Replace (Заменить). Вид диаграммы изменится в соответствии с выбранным стилем.

Что касается меню Chart (Диаграмма), то соответствующие команды перечислены ниже.

- Options (Параметры) — задание параметров, управляющих представлением диаграммы, включая расположение и размеры столбцов для столбиковой диаграммы, положение первого сегмента для круговой диаграммы и т. п.
- Axis (Ось) — определение величин, соответствующих вертикальной и горизонтальной осям, задание масштаба, диапазонов, рисок на осях, вспомогательной сетки, а также содержания и расположения заголовков и меток.
- Bar Spacing (Ширина столбцов) — задание ширины столбцов и расстояния между ними для гистограмм и столбиковых диаграмм.
- Title (Заголовок) — создание и редактирование заголовков и подзаголовков.
- Footnote (Сноска) — создание, редактирование и задание положения сносок.
- Legend (Легенда) — создание и редактирование элементов легенды диаграммы.
- Annotation (Аннотация) — редактирование расположения и содержимого меток осей.
- Reference Line (Идентификационная линия) — открытие диалогового окна для создания одной или нескольких идентификационных линий в диаграмме. В качестве идентификационных линий могут выступать среднее значение, медиана или любая другая величина, упрощающая понимание диаграммы.
- Outer Frames (Внешняя рамка) и Inner Frame (Внутренняя рамка) — создание и удаление внешних и внутренних рамок.
- Refresh (Обновить) — перерисовка диаграммы, необходимая в тех случаях, когда диаграмма редактируется одновременно несколькими графическими инструментами программы.

# 6 Частоты



В этой главе рассматриваются частоты, их графическое представление (столбиковые и круговые диаграммы), гистограммы и процентили. Команда Frequencies (Частоты) относится к числу тех, которые позволяют непосредственно работать с диаграммами, вообще не прибегая к командам меню Graphs (Графики). Некоторые детали, касающиеся редактирования диаграмм, были рассмотрены в главе 5. *Столбиковые* и *круговые диаграммы* отображают количество объектов *(частоту)* в различных категориях дискретной переменной (имеющей небольшое число возможных значений). Однако команда Frequencies (Частоты) позволяет работать не только с дискретными, но и с непрерывными переменными, которые могут принимать большое число разных значений.

Далее даны краткие определения понятий, о которых пойдет речь в этой главе.

- *Частоты.* Команда Frequencies (Частоты) является одной из самых простых и часто используемых команд SPSS. Действие команды сводится к простому подсчету распределения частот по категориям переменной: определению количества объектов в каждой категории переменной. Если мы анализируем переменную пол, то программа подсчитает распределение численности девушек и юношей среди учащихся; если используется переменная класс, то мы получим распределение численности учащихся по классам. Выводимый результат для каждой категории включает метку значения нереМенной, само значение переменной, частоту, процент и накопленный процент от общей частоты.
- *Столбиковые диаграммы.* Обязательное условие для применимости столбиковой диаграммы — дискретность отображаемых данных. Такими свойствами обладают переменные пол, класс, хобби и т. п. Каждая из этих переменных делит все объекты файла данных па категории. Например, переменная вуз разбивает всех учащихся на 4 группы по выбору профиля дальнейшего образования. Непрерывные данные также допускают разбиение па категории, однако эти категории определяются не отдельными значениями переменной, а их диапазонами (интервалами). Примерами непрерывных переменных являются средний балл отметки каждого учащегося, время финиша в гоночных

состязаниях, вес пациента при медосмотре и т. п. Для отображения непрерывных данных используются гистограммы.

- *Гистограммы* используются для отображения распределения частот непрерывных переменных. Особенностью гистограмм, отличающей их от столбиковых диаграмм, является то, что каждый столбец представляет не отдельное значение, а диапазон значений переменной. Гистограммы могут использоваться и для дискретных данных в случаях, если число значений переменной слишком велико, чтобы отображать каждое из них отдельным столбцом диаграммы. Весь диапазон изменчивости переменной разделяют на интервалы, а затем, исходя из этих интервалов, строят гистограмму. Именно таким методом следовало бы отображать распределение частот для переменных отметка1 или отметка2.
- *Процентиль* показывает, какой процент распределения лежит ниже заданной величины. Например, если говорят, что процентиль значения 111 равен 75, то это означает, что 75 % всех значений переменной меньше, чем 111, а 25 % – больше, чем 111. Наиболее часто процентили применяются в педагогических и психологических исследованиях.

# Пошаговые алгоритмы вычислений

Для иллюстрации вычислений мы будем использовать уже знакомый вам файл ex01.sav (мы не приводим примеры круговых диаграмм ввиду их исключительной простоты). Число объектов в файле равно 100.

Первой мы рассмотрим команду Frequencies (Частоты), однако сначала следует выполнить три подготовительных шага.

**Шаг 1** Создайте новый файл данных или подготовьте существующий. Все необходимые действия описаны в главе 3.

**Шаг 2** Запустите программу SPSS через главное меню или при помощи значка на рабочем столе и в открывшемся после запуска программы диалоговом окне SPSS for Windows щелкните на кнопке Cancel (Отмена).

После выполнения этого шага на экране появится окно редактора данных SPSS.

**Шаг 3** Откройте файл данных, с которым вы намерены работать (в пашем случае это файл ex01 .sav). Если он расположен в текущей папке, то выполните следующие действия:

1. Выберите в меню File (Файл) команду Open ▸ Data (Открытие ▸ Данные) или щелкните на кнопке Open File (Открытие файла) панели инструментов.
2. В открывшемся диалоговом окне дважды щелкните на имени ex01.sav или введите его с клавиатуры и щелкните на кнопке OK.

Независимо от того, открыта программа SPSS только что или какие-то процедуры уже выполнялись, в верхней части главного окна должна присутствовать строка меню (она показана ниже). Пока строка меню присутствует на экране, доступны все команды анализа данных. При этом окно с данными видеть не обязательно.




При работе с таблицами результатов или при редактировании диаграмм некоторые пункты строки меню могут исчезать или меняться. Чтобы вернуться к главному окну и стандартной строке меню, щелкните на кнопке свертывания или восстановления текущего окна.

После завершения шага 3 на экране должно присутствовать окно редактора данных со строкой меню.

**Шаг 4** В меню Analyze (Анализ) выберите команду Descriptive Statistics ▸ Frequencies (Описательные статистики ▸ Частоты). На экране появится диалоговое окно, представленное на рис. 6.1.

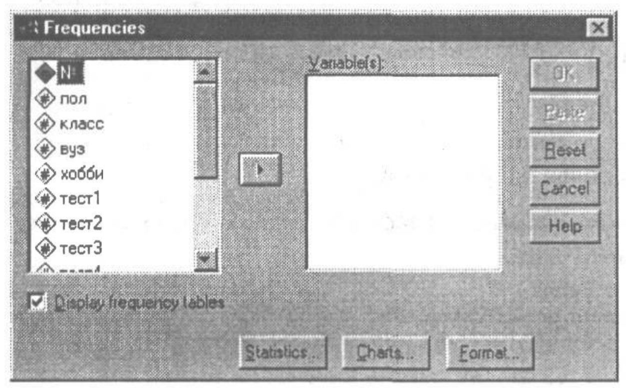



**Рис. 6.1.** Диалоговое окно Frequencies

# Частоты

Диалоговое окно Frequencies (Частоты) типично для большинства статистических операций SPSS. В левой части окна расположен список всех доступных переменных. В этом списке необходимо выбрать те переменные, для которых необходимо вычислить распределение частот.

Для этого щелчком выделите нужную переменную, а затем щелчком на кнопке с направленной вправо стрелкой переместите ее в целевой список Variable(s) (Переменные). Если желаемая переменная не видна в исходном списке, воспользуйтесь полосой прокрутки.

Для того чтобы удалить переменную из целевого списка, достаточно выделить ее в нем, а затем воспользоваться кнопкой с направленной влево стрелкой, которая появляется на месте предыдущей кнопки при выделении любого пункта в списке Variable(s) (Переменные). В этом случае переменная вновь переместится в исходный список. Чтобы полностью очистить список Variable(s) (Переменные), щелкните на кнопке Reset (Сброс).

**Шаг 5** На этом шаге займемся вычислением частот для переменных пол, класс и вуз.
• *1* При открытом диалоговом окис Frequencies (Частоты) выполните слсдующие действия:

1. Щелкните сначала на переменной пол, чтобы выделить се, а затем на кнопке со стрелкой, чтобы переместить переменную в список Variable(s) (Переменные).
2. Повторите предыдущее действие для переменных класс и вуз.
3. Щелкните па кнопке OK, чтобы открыть окно вывода, показанное па рис. 6.2.

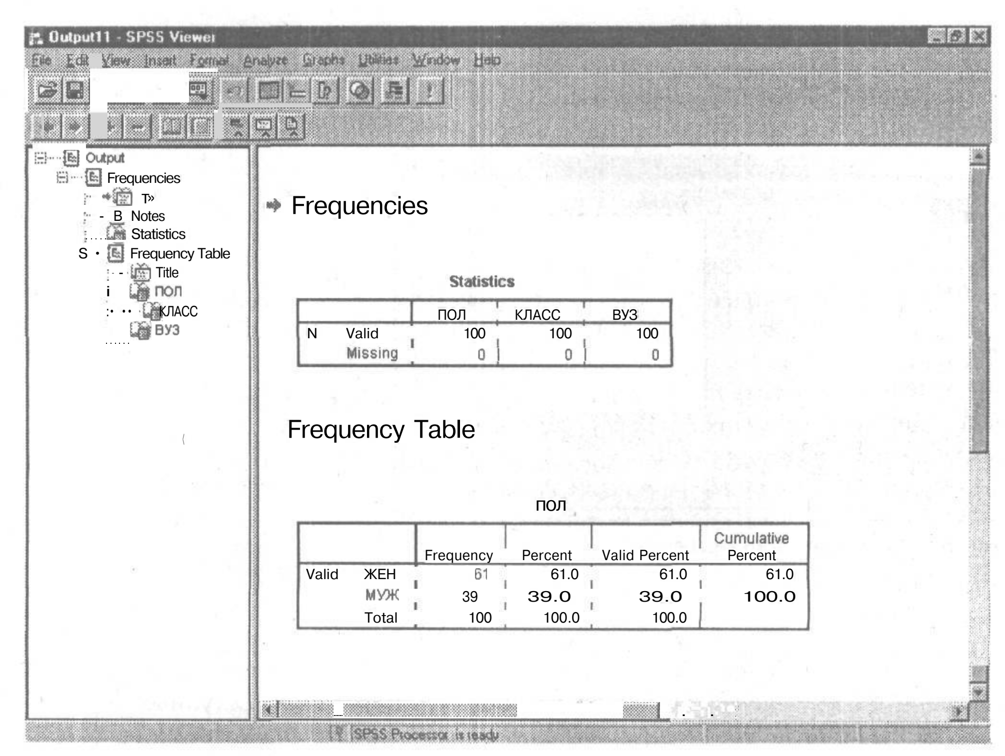



**Рис. 6.2.** Окно вывода программы SPSS

Таким образом, переместив переменные в целевой список, вы завершаете выполнение операции щелчком на кнопке OK. После этого программа SPSS формирует окно вывода с результатами выполнения команды. Выполнение практически всех команд статистических операций завершается именно так.

Окно с результатами работы программы имеет заголовок Output1 — SPSS Viewer (Вывод1 — Просмотр SPSS). Для просмотра результатов при необходимости можно воспользоваться полосой прокрутки в правой части окна. Структура выводимых результатов программы и основные приемы их редактирования описаны в разделе «Окно вывода» главы 2. Вы можете снова применить команду Frequencies (Частоты) к выбранному подмножеству данных: для этого достаточно выбрать команду Analyze ▸ Descriptive Statistics ▸ Frequencies (Анализ ▸ Описательные статистики ▸ Частоты) и в открывшемся диалоговом окне сразу щелкнуть на кнопке OK (ранее выбранные переменные окажутся в целевом списке автоматически) или предварительно изменить список переменных для анализа.

## Столбиковые диаграммы

Для того чтобы создать столбиковую диаграмму для дискретных данных, необходимо сначала выполнить все действия шага 5, кроме последнего, а затем в диалоговом окне Frequencies (Частоты) щелкнуть на кнопке Charts (Диаграммы). На экране появится диалоговое окно Frequencies: Charts (Частоты: Диаграммы), представленное на рис. 6.3.

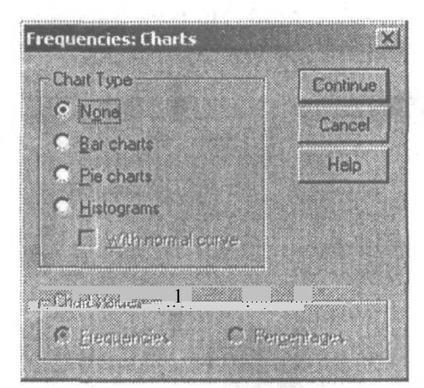



**Рис. 6.3.** Диалоговое окно Frequencies: Charts

Как видите, диалоговое окно Frequencies: Charts (Частоты: Диаграммы) позволяет выбрать тип диаграммы с помощью переключателей Bar charts (Столбиковая), Pie charts (Круговая) и Histograms (Гистограмма). В данном примере вам понадобится установить переключатель Bar charts (Столбиковая). В зависимости от величины, которую вы хотите использовать для отображения частот, в группе Chart Values (Значения на диаграмме) установите переключатель Frequencies (Частоты) или Percentages (Проценты). Для закрытия диалогового окна Frequencies: Charts (Частоты: Диаграммы) щелкните на кнопке Continue (Продолжить). После этого вы вернетесь в диалоговое окно процедуры Frequencies (Частоты) и сможете завершить операцию щелчком на кнопке OK.

**Шаг 5а** На этом шаге мы выполним построение столбиковых диаграмм для частот тех же трех переменных пол, класс и вуз, которые использовались ранее. После выполнения шага 4 у вас должно быть открыто диалоговое окно Frequencies (Частоты), показанное на рис. 6.1. Для разнообразия целевой список мы будем заполнять двойными щелчками, а не щелчками на кнопке со стрелкой, как в предыдущем примере.

1. Дважды щелкните на переменной пол. Ее имя появится в списке Variable(s) (Переменные). Сделайте то же самое для переменных класс и вуз и щелкните на кнопке Charts (Диаграммы), чтобы открыть диалоговое окно Frequencies: Charts (Частоты: Диаграммы).
2. В группе Chart Type (Тип диаграммы) установите переключатель Bar charts (Столбиковая). Если вы предпочитаете вместо частот выводить на диаграмме их доли от общей частоты в процентах, в группе Chart Values (Значения на диаграмме) установите переключатель Percentages (Проценты) и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Frequencies (Частоты).
3. Щелкните на кнопке ОК, чтобы открыть окно вывода.

После выполнения этого шага программа сгенерирует три диаграммы, соответствующие выбранным переменным. При помощи полосы прокрутки вы можете просмотреть созданные диаграммы в окне вывода. Если у вас возникнет необходимость отредактировать какую-либо из диаграмм, дважды щелкните на ней и обратитесь к материалу главы 5 для дальнейших действий. На рис. 6.4 представлено окно вывода с диаграммой частот для переменной вуз.

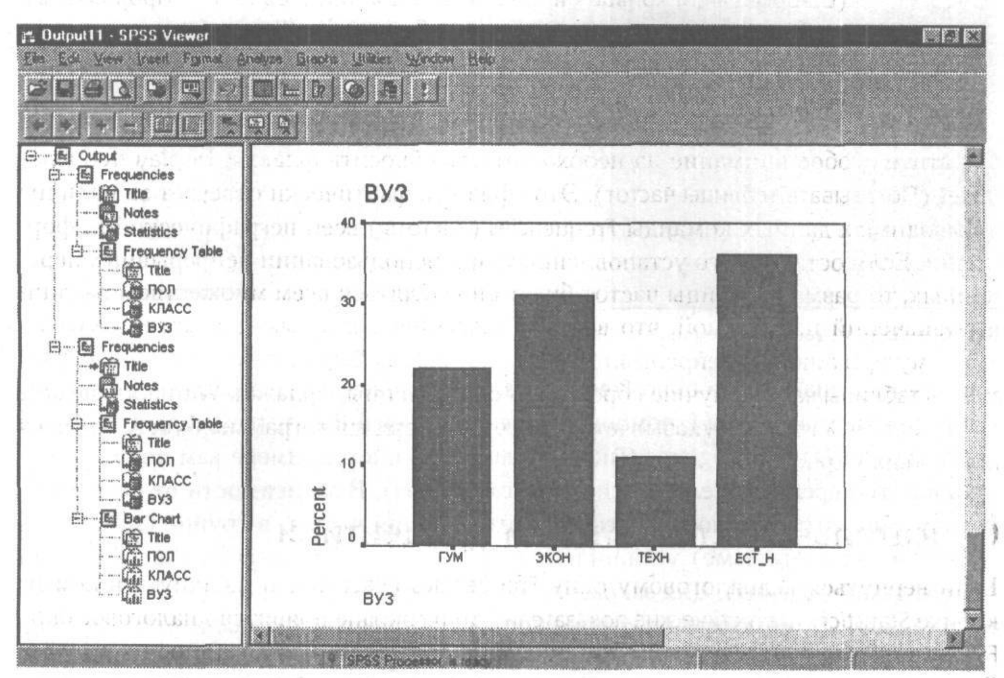



**Рис. 6.4.** Пример столбиковой диаграммы

## Гистограммы

Процесс построения гистограмм аналогичен процессу построения столбиковых диаграмм. Единственным отличием, которое следует всегда иметь в виду, является то, что гистограммы предназначены для отображения распределения непрерывных переменных. Таким образом, для переменных пол, класс и вуз гистограммы непригодны. В то же время для переменной отметка2, отражающей среднюю отметку для каждого учащегося, гистограмма является весьма удобным средством описания распределения частот. Ниже приведен пример построения гистограммы для переменной отметка2.

**Шаг 5б** После выполнения шага 4 у вас должно быть открыто диалоговое окно Frequencies (Частоты), показанное на рис. 6.1. При необходимости повторите шаги 1-4. Если у вас в целевом списке диалогового окна Frequencies (Частоты) после работы над материалом предыдущего раздела остались переменные, щелкните на кнопке Reset (Сброс).

1. Щелкните сначала на переменной отметка2, чтобы выделить ее, затем на кнопке со стрелкой, чтобы переместить переменную в список Variable(s) (Переменные), и, наконец, на кнопке Charts (Диаграммы), чтобы открыть диалоговое окно Frequencies: Charts (Частоты: Диаграммы).
2. В группе Chart Type (Тип диаграммы) установите переключатель Histograms (Гистограммы); если необходимо, установите флажок With normal curve (С нормальной кривой) и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Frequencies (Частоты).
3. В диалоговом окне Frequencies (Частоты) сбросьте флажок Display frequency tables (Показывать таблицы частот) и щелкните на кнопке OK.

Обратите особое внимание на необходимость сбросить флажок Display frequency tables (Показывать таблицы частот). Этот флажок фактически отвечает за наличие в выводимых данных команды Frequencies (Частоты) всей неграфической информации. Если оставить его установленным при использовании непрерывных переменных, то размер таблицы частот будет определяться всем множеством различных значений переменной, что вряд ли окажется полезным для исследователя. Поэтому при анализе непрерывных данных флажок Display frequency tables (Показывать таблицы частот) лучше сбросить. Установленный флажок With normal curve (С нормальной кривой) указывает на то, что поверх гистограммы будет нанесена нормальная кривая.

## Описательные статистики и процентили

Если вернуться к диалоговому окну Frequencies (Частоты) и щелкнуть в нем на кнопке Statistics (Статистические показатели), то на экране появится диалоговое окно Frequencies: Statistics (Частоты: Статистические показатели), показанное на рис. 6.5. В этом окне представлены такие показатели, как процентили и описательные

статистики (подробно описательные статистики рассматриваются в главе 7). Из трех приведенных далее примеров вы узнаете, каким образом создать гистограмму, получить описательную статистическую информацию и вычислить процентили с заданным шагом или заданной величины. Во всех трех примерах используется переменная отметка2, а начинаются они с диалогового окна, полученного после выполнения шага 4 (см. рис. 6.1).

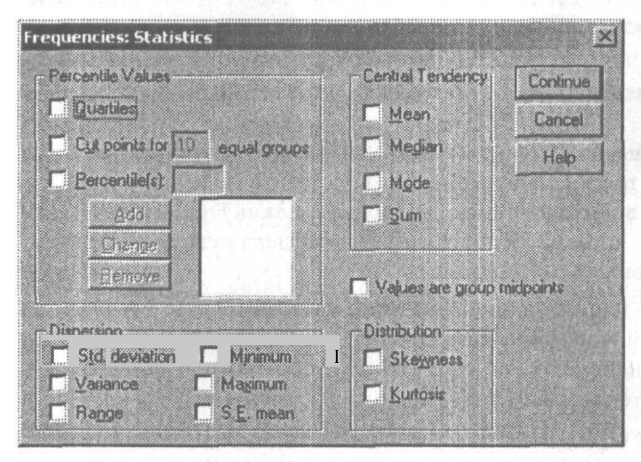



**Рис. 6.5.** Диалоговое окно Frequencies: Statistics

**Шаг 5в** На этом шаге мы построим гистограмму распределения частот и вычислим для него среднее значение, стандартное отклонение, асимметрию и эксцесс распределения.

1. Щелкните сначала на переменной отметка2, чтобы выделить ее, затем — на кнопке со стрелкой, чтобы переместить переменную в список Variable(s) (Переменные), и, наконец, — на кнопке Charts (Диаграммы), чтобы открыть диалоговое окно Frequencies: Charts (Частоты: Диаграммы).
2. В группе Chart Type (Тип диаграммы) установите переключатель Histograms (Гистограмма) и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Frequencies (Частоты).
3. Щелкните на кнопке Statistics (Статистические показатели), чтобы открыть диалоговое окно Frequencies: Statistics (Частоты: Статистические показатели), показанное на рис. 6.5.
4. В группе Central Tendency (Показатели центральной тенденции) установите флажок Mean (Среднее), в группе Distribution (Распределение) — флажки Skewness (Асимметрия) и Kurtosis (Эксцесс), в группе Dispersion (Изменчивость) — флажок Std. deviation (Стандартное отклонение) и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Frequencies (Частоты).
5. В диалоговом окне Frequencies (Частоты) сбросьте флажок Display frequency tables (Показывать таблицы частот) и щелкните на кнопке OK.

Шаг 5г Чтобы вычислить процентили переменной отметка2 с шагом 10, необходимо разделить шкалу процентилей на 10 равных промежутков с границами, равными 10, 20, 30 и т. д.

1. Щелкните сначала на переменной отметка2, чтобы выделить ее, затем — на кнопке со стрелкой, чтобы переместить переменную в список Variable(s) (Переменные), и, наконец, на кнопке Statistics (Статистические показатели), чтобы открыть диалоговое окно Frequencies: Statistics (Частоты: Статистические показатели), показанное на рис. 6.5.
2. В группе Percentile Values (Значения процентилей) установите флажок Cut points for ... equal groups (Разделить точки на ... равных групп), в поле рядом с флажком введите значение 20 и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Frequencies (Частоты).
3. В диалоговом окне Frequencies (Частоты) сбросьте флажок Display frequency tables (Показывать таблицы частот) и щелкните на кнопке OK.

Шаг 5д Наконец, чтобы вычислить набор произвольных процентилей (в данном случае — 5, 70 и 95), выполните следующие действия:

1. Щелкните сначала на переменной отметка2, чтобы выделить ее, затем — на кнопке со стрелкой, чтобы переместить переменную в список Variable(s) (Переменные), и, наконец, на кнопке Statistics (Статистические показатели), чтобы открыть диалоговое окно Frequencies: Statistics (Частоты: Статистические показатели), показанное на рис. 6.5.
2. В группе Percentile Values (Значения процентилей) установите флажок Percentiles (Процентили), в поле рядом с флажком введите значение 5 и щелкните на кнопке Add (Добавить).
3. Повторите предыдущее действие для значений 70 и 95 и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Frequencies (Частоты).
4. В диалоговом окне Frequencies (Частоты) сбросьте флажок Display frequency tables (Показывать таблицы частот) и щелкните на кнопке OK.

Обратите внимание на то, что квартили (25-й, 50-й и 75-й процентили) можно вычислить одним действием, установив флажок Quartiles (Квартили) в группе Percentile Values (Значения процентилей) диалогового окна Frequencies: Statistics (Частоты: Статистические показатели).

# Печать результатов и выход из программы

Ниже описана типичная процедура печати результатов применения статистического метода (или нескольких методов). После выполнения шага 5 должно быть открыто окно вывода результатов.

Шаг 6 В окне вывода укажите фрагменты, выводимые на печать (см. раздел «Окно вывода» в главе 2), в меню File (Файл) выберите команду Print (Печать), при необходимости задайте параметры печати и щелкните на кнопке OK.

Последнее, что необходимо сделать после завершения обработки данных и печати результатов, — это выйти из программы SPSS.

Шаг 7 Для выхода из программы в меню File (Файл) выберите команду Exit (Выход).

Иногда после выполнения команды Exit (Выход) на экране могут появляться небольшие диалоговые окна с вопросом о необходимости сохранения сделанных в файлах изменений и кнопками, описывающими возможные варианты ответа. Для завершения работы просто щелкайте на соответствующих кнопках.

# Представление результатов

Несмотря на то что выводимые программой SPSS результаты занимают не очень много места, как правило, их представляют в еще более компактной форме (по вопросам редактирования выводимых результатов см. раздел «Окно вывода» в главе 2).

На возможные небольшие различия между данными, которые вы видите на экране своего компьютера, и иллюстрациями этой книги внимания можно не обращать.

## Частоты

На рис. 6.6 приведен фрагмент окна вывода после выполнения шага 5.

Ниже дана трактовка терминов, используемых программой в окне вывода.

- Frequency (Частота) — число объектов, соответствующих каждой категории (градации) переменной.
- Percent (Процент) — процент от общей численности (с учетом пропусков). Если бы в файле были пропущенные значения, то их процент был бы указан в предпоследней строке.
- Valid percent (Действительный процент) — процент значений для каждой категории за вычетом пропущенных значений.
- Cumulative percent (Накопленный процент) — накопленный процент величины Valid percent (Действительный процент).
- Valid (Действительное значение) — список уровней переменной.
- Total (Всего) — итоговые значения.

ПОЛ

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | ЖЕН | 61 | 61.0 | 61.0 | 61.0 |
| | МУЖ | 39 | 39.0 | 39.0 | 100.0 |
| | Total | 100 | 100.0 | 100.0 | |

КЛАСС

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | 1 | 33 | 33.0 | 33.0 | 33.0 |
| | 2 | 35 | 35.0 | 35.0 | 68.0 |
| | 3 | 32 | 32.0 | 32.0 | 100.0 |
| | Total | 100 | 100.0 | 100.0 | |

ВУЗ

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | ГУМ | 22 | 22.0 | 22.0 | 22.0 |
| | ЭКОН | 31 | 31.0 | 31.0 | 53.0 |
| | ТЕХН | 11 | 11.0 | 11.0 | 64.0 |
| | ЕСТ_Н | 36 | 36.0 | 36.0 | 100.0 |
| | Total | 100 | 100.0 | 100.0 | |

**Рис. *6.6.*** Фрагмент окна вывода после выполнения шага 5

## Гистограммы

На рис. 6.7 представлены результаты выполнения шага 5б. По горизонтальной оси отложены значения переменной с шагом, равным 0,125 (ширина каждого столбца тоже равна 0,125). Каждое указанное значение соответствует середине столбца. На вертикальной оси гистофаммы отложены частоты диапазонов. Справа от гистофаммы помещены вычисленные среднее значение (Mean) и стандартное отклонение распределения (Std. Dev), а также общее число объектов (N). Кроме того, на гистофамму нанесена нормальная кривая, поскольку при выполнении процедуры был установлен флажок With normal curve (С нормальной кривой). Гистограмма без труда может быть сделана более наглядной. В частности, для данной гистофаммы желательно укрупнение интервалов и, соответственно, уменьшение количества столбцов. По вопросам редактирования гистофамм обратитесь к главе 5.

## Описательные статистики и процентили

На рис. 6.8 представлены результаты выполнения шагов 5г и 5д.

Описательные статистики рассматриваются подробнее в главе 7, поэтому мы не станем обсуждать их здесь. Обратите внимание, что если в вывод результатов включены асимметрия (Skewness) и эксцесс (Kurtosis), то для них по умолчанию вычисляется стандартная ошибка (Std. Error of Skewness и Std. Error of Kurtosis соответственно).

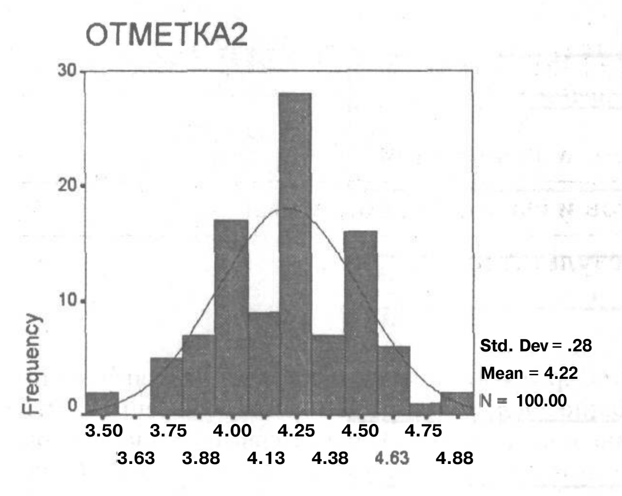



**Рис. 6.7.** Фрагмент окна вывода после выполнения шага 5б

ОТМЕТКА2

| | | |
|---|---|---|
| N | Valid | 100 |
| | Missing | 0 |
| Std. Deviation | | .27589 |
| Skewness | | -.007 |
| Std. Error of Skewness | | .241 |
| Kurtosis | | -.357 |
| Std. Error of Kurtosis | | .478 |

ОТМЕТКА2

| | | |
|---|---|---|
| N | Valid | 100 |
| | Missing | 0 |
| Percentiles | 10 | 3.8550 |
| | 20 | 3.9500 |
| | 30 | 4.0500 |
| | 40 | 4.1700 |
| | 50 | 4.2000 |
| | 60 | 4.3000 |
| | 70 | 4.3500 |
| | 80 | 4.4900 |
| | 90 | 4.5500 |

**Рис. 6.8.** Фрагменты окна вывода после выполнения шагов 5г и 5д

Что касается процентилей (Percentiles), то их можно трактовать следующим образом: для переменной отметка2 10 % значений не превышают 3,855 (10 % учащихся имеют отметку не выше 3,855; 30 % значений не превышают 4,05 (30 % учащихся имеют отметку не выше 4,05), и т. д.

# 7 Описательные статистики



*Описательные статистики* (descriptive statistics) — это различные вычисляемые показатели, характеризующие распределение значений переменной. Эти показатели условно можно разбить на несколько групп. Первая группа — меры центральной тенденции, вокруг которых «группируются» данные: среднее значение, медиана и мода. Вторая группа характеризует изменчивость значений переменной относительно среднего: стандартное отклонение и дисперсия. Диапазон изменчивости характеризуется минимумом, максимумом и размахом. Асимметрия и эксцесс представляют меру отклонения формы распределения от нормального вида. Кроме того, существуют величины, выражающие погрешности некоторых статистик: стандартная ошибка среднего, стандартная ошибка асимметрии и стандартная ошибка эксцесса. Последние два показателя вычисляются программой вместе с асимметрией и эксцессом по умолчанию. При помощи команды Descriptives (Описательные статистики) можно вычислить любую из указанных величин.

## Меры центральной тенденции

Существует три основные меры центральной тенденции распределения.

- *Среднее значение* (mean) равно сумме всех значений распределения, деленной па их количество. Для распределения [3 5 7 5 6 8 9] среднее значение равно (3 + 5 + 7 + 5 + 6 + 8 + 9)/7 = 6,14.
- *Медиана* (median) определяется как значение, находящееся в середине распределения, полученного из исходного путем упорядочивания по возрастанию. Для распределения [3 5 7 5 6 8 9J медиана равна 6, поскольку значение, равное 6, находится в центре последовательности [3 5 5 6 7 8 9].
- *Мода* (mode) равна наиболее часто встречающемуся значению. В распределении [3 5 7 5 6 8 9] мода равна 5, поскольку число 5 встречается в нем дважды.

## Меры изменчивости

Выделяют две величины, характеризующие изменчивость, или разброс, значений распределения относительно среднего.

- *Дисперсия* (variance) равна сумме квадратов отклонений каждого значения от среднего, деленной на $N - 1$, где $N$ — число значений в распределении. Для распределения [3 5 7 5 6 8 9] дисперсия равна $((3 - 6{,}14)^2 + (5 - 6{,}14)^2 + (7 - 6{,}14)^2 + (5 - 6{,}14)^2 + (6 - 6{,}14)^2 + (8 - 6{,}14)^2 + (9 - 6{,}14)^2)/6 - 4{,}1429$.
- *Стандартное отклонение* (standard deviation) равно квадратному корню из дисперсии. Для распределения [3 5 7 5 6 8 9] стандартное отклонение равно 2,0354.

Стандартное отклонение является довольно наглядной и информативной для исследователя характеристикой распределения, а дисперсия, как правило, используется как вспомогательная величина в статистических вычислениях.

## Характеристики диапазона распределения

Дополнительными мерами изменчивости являются 4 простые характеристики, отражающие границы распределения и его размах.

*Минимум* (minimum) равен наименьшему из значений распределения. Для распределения [3 5 7 5 6 8 9] минимум равен 3.

*Максимум* (maximum) равен наибольшему из значений распределения. Для распределения [3 5 7 5 6 8 9] максимум равен 9.

*Размах* (range) составляет разность между максимумом и минимумом распределения. В случае распределения [3 5 7 5 6 8 9] размах равен $9 - 3 = 6$.

*Сумма* (sum) равна сумме всех значений распределения. Для распределения [3 5 7 5 6 8 9] сумма равна $3 + 5 + 7 + 5 + 6 + 8 + 9 = 43$.

## Характеристики формы распределения

Для отражения близости формы распределения к нормальному виду существует две основные характеристики.

- *Эксцесс* (kurtosis) является мерой «сглаженности» («остро-» или «плосковершинности») распределения. Если значение эксцесса близко к 0, это означает, что форма распределения близка к нормальному виду. Положительный эксцесс указывает на «плосковершинное» распределение, у которого максимум вероятности выражен не столь ярко, как у нормального. Значения эксцесса, превышающие 5,0, говорят о том, что по краям распределения находится больше значений, чем вокруг среднего. Отрицательный эксцесс, напротив, характеризует «островершинное» распределение, график которого более вытянут по вертикальной оси, чем график нормального распределения. Считается, что распределение с эксцессом в диапазоне от -1 до +1 примерно соответствует нормальному виду. В большинстве случаев вполне допустимо считать нормальным распределение с эксцессом, по модулю не превосходящим 2.

- *Асимметрия* (skewness) показывает, в какую сторону относительно среднего сдвинуто большинство значений распределения. Нулевое значение асимметрии означает симметричность распределения относительно среднего значения, положительная асимметрия указывает па сдвиг распределения в сторону меньших значений, а отрицательная асимметрия — в сторону больших значений. В большинстве случаев за нормальное принимается распределение с асимметрией, лежащей в пределах от -1 до +1. В исследованиях, не требующих высокой точности результатов, нормальным считают распределение с асимметрией, по модулю не превосходящей 2.

## Стандартная ошибка

*Стандартная ошибка* (standard error) является характеристикой точности, или стабильности, величины, для которой она вычисляется. В контексте программы SPSS стандартная ошибка используется для среднего значения, асимметрии и эксцесса. Ее смысл заключается в следующем. Вы можете, взяв определенное количество случайно выбранных значений генеральной совокупности, составить выборку и вычислить для нее среднее значение. Повторив эту операцию некоторое количество раз, вы получите набор средних значений выборок, которые также представляют собой некоторое распределение. Стандартное отклонение этого распределения и будет являться стандартной ошибкой для среднего значения генеральной совокупности. Аналогичным способом вычисляются стандартные ошибки для асимметрии и эксцесса. Чем меньше значение стандартной ошибки, тем выше стабильность величины, для которой она вычисляется.

# Пошаговый алгоритм вычислений

Для применения команды Descriptives (Описательные статистики) мы обратимся к файлу ex01 .sav. Число объектов, или значений каждой переменной, в этом файле равно 100, поэтому при вычислении характеристик распределения для различных переменных программа будет считать *N* равным 100. Сначала необходимо выполнить три подготовительных шага.

Шаг 1 Создайте новый файл данных или подготовьте существующий. Все необходимые действия описаны в главе 3.

Шаг 2 Запустите программу SPSS при помощи значка на рабочем столе или команды Пуск ▸ Программы ▸ SPSS for Windows ▸ SPSS 11.5 for Windows (Start ▸ Programs ▸ SPSS for Windows ▸ SPSS 11.5 for Windows) и в открывшемся после запуска программы диалоговом окне SPSS for Windows щелкните на кнопке Cancel (Отмена).

После выполнения этого шага на экране появится окно редактора данных SPSS.

Шаг 3 | Откройте файл данных, с которым вы намерены работать (в пашем случае это файл ex01 .sav). Если он расположен в текущей папке, то выполните следующие действия:

1. Выберите в меню File (Файл) команду Open ▸ Data (Открытие ▸ Данные) или щелкните на кнопке Open File (Открытие файла) панели инструментов.
2. В открывшемся диалоговом окне дважды щелкните на имени ex01.sav или введите его с клавиатуры и щелкните па кнопке OK.

Независимо от того, открыта программа SPSS только что или какие-то процедуры уже выполнялись, в верхней части главного окна должна присутствовать строка меню (она показана ниже). Пока строка меню присутствует на экране, доступны все команды анализа данных. При этом окно с данными видеть не обязательно.




При работе с таблицами результатов или при редактировании диаграмм некоторые пункты строки меню могут исчезать или меняться. Чтобы вернуться к главному окну и стандартной строке меню, щелкните на кнопке свертывания или восстановления текущего окна.

После завершения шага 3 на экране должно присутствовать окно редактора данных со строкой меню.

Шаг 4 В меню Analyze (Анализ) выберите команду Descriptive Statistics ▸ Descriptives (Описательные статистики ▸ Описательные статистики). На экране появится диалоговое окно Descriptives (Описательные статистики), показанное на рис. 7.1.

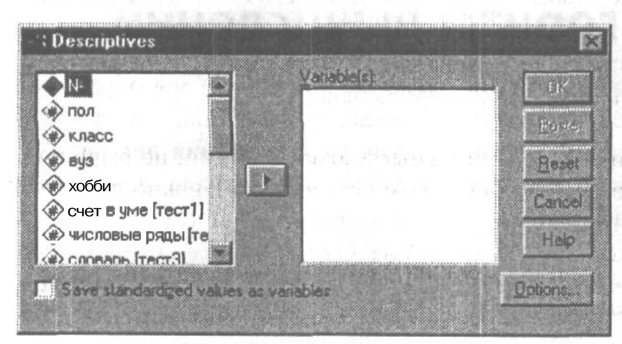



Рис. 7.1. Диалоговое окно Descriptives

В диалоговом окне Descriptives (Описательные статистики) необходимо задать переменные, для которых будут вычислены описательные статистики. В левой части окна находится список всех доступных переменных текущего файла данных.

Для задания переменной щелчком выделите нужную переменную, а затем щелчком на кнопке с направленной вправо стрелкой переместите ее



в целевой список Variable(s) (Переменные). Если желаемая переменная не видна в исходном списке, воспользуйтесь полосой прокрутки.

Для того чтобы удалить переменную из целевого списка, достаточно выделить ее в нем, а затем воспользоваться кнопкой с направленной влево стрелкой, которая появляется на месте предыдущей кнопки при выделении любого пункта в списке Variable(s) (Переменные). В этом случае переменная вновь переместится в исходный список. Чтобы полностью очистить список Variable(s) (Переменные), щелкните на кнопке Reset (Сброс).

Как вы могли заметить, помимо кнопок и списков диалоговое окно Descriptives (Описательные статистики) содержит флажок Save standardized values as variables (Сохранять стандартизованные значения как переменные). Если этот флажок установлен, то программа произведет *z*-преобразование (стандартизацию) всех выбранных переменных, создав таким образом новые переменные. Исходные переменные при этом останутся без изменений, а новым переменным будут присвоены старые имена, но начинающиеся с буквы z. Под стандартизованными (или z-преобразованными) значениями переменной понимается такое ее распределение, среднее значение которого равно 0, а стандартное отклонение — 1. Обратите внимание, что в слове «standardized» названия флажка буква «z» подчеркнута. Это должно напоминать о том, что эта буква является первой (префиксом) в именах стандартизованных переменных.

**Шаг 5** Для того чтобы создать таблицу описательных статистик, предлагаемую программой по умолчанию и включающую среднее значение, стандартное отклонение, максимум и минимум, выполните следующие действия:

1. Щелкните сначала на переменной отметка1, чтобы выделить ее, а затем на кнопке со стрелкой, чтобы переместить переменную в список Variable(s) (Переменные).
2. Повторите те же действия для переменной отметка2.
3. Щелкните на кнопке OK, чтобы открыть окно вывода.

Если вам понадобится вычислить дополнительные характеристики, не вычисляемые программой по умолчанию, то перед щелчком на кнопке OK нужно щелкнуть на кнопке Options (Параметры), расположенной в нижнем правом углу диалогового окна Descriptives (Описательные статистики). Откроется диалоговое окно Descriptives: Options (Описательные статистики: Параметры), в котором с помощью флажков можно задать все упоминавшиеся выше характеристики, за исключением двух: медианы и моды (рис. 7.2). Последние две характеристики доступны только через команду Frequencies (Частоты), о которой рассказывалось в главе 6. Кроме того, вы не увидите в диалоговом окне флажков, соответствующих стандартным ошибкам асимметрии и эксцесса, поскольку они всегда вычисляются автоматически. Установите флажки, соответствующие нужным характеристикам, и щелкните на кнопке Continue (Продолжить). При желании можно также установить один из четырех переключателей в группе Display Order (Порядок отображения). Переключатель Variable list (Список переменных) установлен по умолчанию

и означает, что в выводимых результатах переменные будут перечислены в том же порядке, в котором они представлены в файле данных. Переключатель Alphabetic (Алфавитный) указывает, что перечисление переменных будет происходить в алфавитном порядке. Переключатели Ascending means (По возрастанию средних) и Descending means (По убыванию средних) позволяют отсортировать переменные в выводимых результатах по вычисленным средним значениям.

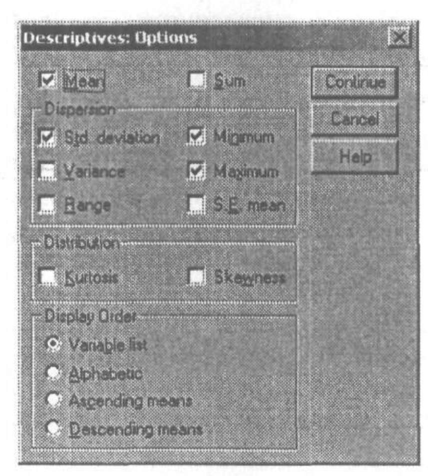



**Рис. 7.2.** Диалоговое окно Descriptives: Options

Далее показано, каким образом можно организовать подсчет произвольных характеристик для переменных отметка1 и отметка2.

Шаг 5а После выполнения шага 4 у вас должно быть открыто окно Descriptives (Описательные статистики), представленное на рис. 7.1.

1. Щелкните сначала на переменной отметка1, чтобы выделить ее, а затем — на кнопке со стрелкой, чтобы переместить переменную в список Variable(s) (Переменные).
2. Повторите те же действия для переменной отметка2 и щелкните на кнопке Options (Параметры), чтобы открыть диалоговое окно Descriptives: Options (Описательные статистики: Параметры).
3. Установите флажки для всех характеристик, которые хотите вычислить, и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Descriptives (Описательные статистики).
4. Щелкните на кнопке OK, чтобы открыть окно вывода, показанное рис. 7.3.

Для того чтобы полностью развернуть окно вывода, воспользуйтесь кнопкой развертывания. Чтобы просматривать результаты внутри окна вывода, используйте вертикальную и горизонтальную полосы прокрутки. Как видите, в верхней части окна вывода имеется строка меню. Это означает, что вы можете выполнять другие статистические процедуры, не обращаясь к окну редактора данных.

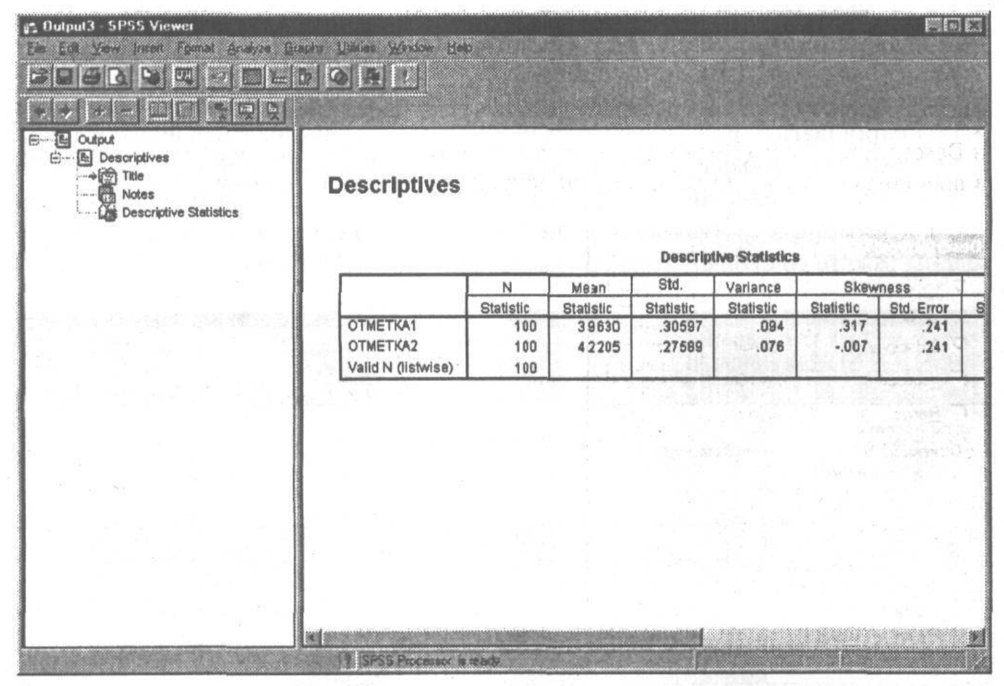



**Рис. 7.3.** Окно вывода SPSS

# Печать результатов и выход из программы

Ниже описана типичная процедура печати результатов обработки (или серии обработок) данных. После выполнения шага 5 должно быть открыто окно вывода.

**Шаг 6** В окне вывода укажите фрагменты, выводимые на печать (см. раздел «Окно вывода» в главе 2), в меню File (Файл) выберите команду Print (Печать), при необходимости задайте параметры печати и щелкните на кнопке OK.

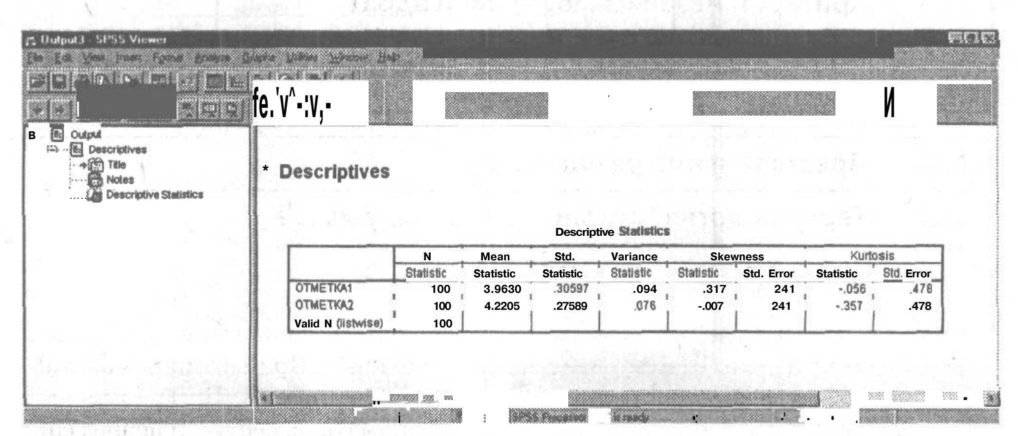
Последнее, что необходимо сделать после завершения исследования и печати результатов, — это выйти из программы SPSS.

**Шаг 7** Для выхода из программы в меню File (Файл) выберите команду Exit (Выход).

Иногда после выполнения команды Exit (Выход) на экране могут появляться небольшие диалоговые окна с вопросом о необходимости сохранения сделанных в файлах изменений и кнопками, описывающими возможные варианты ответа. Для завершения работы просто щелкайте на соответствующих кнопках.

# Представление результатов

На рис. 7.4 приведен один из возможных результатов работы программы после выполнения шага 5. Как видите, в отличие от данных, выводимых после шага 5а, в этот результат добавлена дисперсия (Variance). Помимо дисперсии в результат включено число значений (N), среднее значение (Mean), стандартное отклонение (Std.), асимметрия (Skewness) и эксцесс (Kurtosis). Последние две величины приведены вместе со своими стандартными ошибками (Std Error).




Descriptive Statistics

| | N | Mean | Std. | Variance | Skewness | | Kurtosis | |
|---|---|---|---|---|---|---|---|---|
| | Statistic | Statistic | Statistic | Statistic | Statistic | Std. Error | Statistic | Std. Error |
| OTMETKA1 | 100 | 3.9630 | .30597 | .094 | .317 | 241 | -.056 | .478 |
| OTMETKA2 | 100 | 4.2205 | .27589 | .076 | -.007 | 241 | -.357 | .478 |
| Valid N (listwise) | 100 | | | | | | | |

**Рис. 7.4.** Окно вывода после выполнения шага 5

Обратите внимание, что все выводимые данные уместились на одной странице и целиком видны на экране. Такая ситуация встречается достаточно редко, чаще вам придется пользоваться обеими полосами прокрутки, чтобы видеть результаты анализа. Кроме того, для быстрой навигации внутри окна вывода вы можете использовать иерархическую структуру в виде дерева объектов в левой части окна.

Результаты обработки говорят о том, что в отношении рассматриваемых переменных доступны любые методы статистического анализа: значения асимметрии и эксцесса по модулю не превышают 1 для всех переменных. С терминологией мы уже познакомились (см. раздел «Представление результатов» в главе 6). Единственным исключением является слово «listwise» в последней строке таблицы. Оно указывает на то, что анализ проводился без учета тех объектов (строк), для которых пропущено хотя бы одно значение. Поскольку таких объектов в файле ex01.sav нет, число включенных в анализ объектов (100) равно числу исходных объектов.

# 8 Таблицы сопряженности и критерий хи-квадрат



*Таблицы сопряженности*, или *кросстабуляции*, служат для описания связи двух или более номинативных (категориальных) переменных. Примерами номинативных переменных являются пол (женский, мужской), класс (А, Б, В), местность (город, пригород, село), ответ (да, нет) и т. д. Таблицы сопряженности неприменимы к непрерывным переменным, однако последние можно разбить на интервалы. Так, возраст человека, который следует считать непрерывным из-за большого числа его возможных значений, можно разбить на интервалы от 0 до 19 лет, от 20 до 39 лет, от 40 до 59 лет и т. д. В частности, представление непрерывной переменной в виде интервалов с помощью таблиц сопряженности иногда полезно для их наглядного представления. Напротив, для статистического анализа перевод непрерывных (количественных) переменных в номинативные не целесообразен, так как теряется существенная часть информации о различии объектов. Так, когда два человека в возрасте 39 и 40 лет попадают в соседние возрастные категории, с точки зрения анализа они ничем не будут отличаться от пары людей в возрасте 20 и 59 лет.

Для работы с таблицами сопряженности в программе SPSS используется команда Crosstabs (Таблицы сопряженности).

# Таблицы сопряженности

Обратимся к файлу ex01.sav. С помощью команды Frequencies (Частоты) мы можем узнать, что среди школьников 39 юношей и 61 девушка, что 33 из них увлекаются спортом, 37 — компьютером и 30 — искусством. Однако команда Frequencies (Частоты) не позволяет ответить на вопросы, сколько девушек увлекаются спортом или сколько юношей — искусством. Для этого в SPSS существует команда Crosstabs (Таблицы сопряженности). Вполне логично, что для ответа на наш вопрос

необходимо «сопрячь», или «пересечь», подмножество учащихся определенного пола с подмножеством учащихся с определенным увлечением. Такое сопряжение удобно представить в виде таблицы, строки которой соответствуют полу, столбцы — увлечению. Тогда в ячейке, находящейся, например, на пересечении строки «мужской» и столбца «искусство», мы увидим количество (частоту) юношей, которые увлекаются искусством. Поскольку существуют 2 градации пола и 3 градации внешкольных увлечений (хобби), наша перекрестная таблица будет состоять из 2 х 3 = 6 ячеек. Можно составлять и сложные таблицы сопряженности, включающие три и более переменные, однако эта операция имеет смысл лишь для больших объемов данных, поскольку в противном случае частоты большинства ячеек будут малыми или нулевыми. Рассмотрим, что произойдет, например, если для данных файла ex01 .sav создать таблицу сопряженности пол×хобби×класс×вуз. Эта таблица будет содержать 2 × 3 × 3 × 4 = 72 ячейки; вспомним, что при этом число объектов составляет лишь 100. Очевидно, что большинство ячеек таблицы сопряженности будет иметь значения от 0 до 1-2. При задании этих четырех номинативных переменных программа SPSS вместо «четырехмерной» таблицы построит 12 двухмерных таблиц размерностью 2 × 3, «вложенных» в одну таблицу.

# Критерий независимости хи-квадрат

Помимо частот (или *наблюдаемых* величин) SPSS может вычислять *ожидаемые* значения для каждой ячейки таблицы. Ожидаемое значение вычисляется в предположении, что две номинативные переменные независимы друг от друга. Рассмотрим простой пример. Пусть в комнате находится 100 человек, из которых 30 являются мужчинами, а 70 — женщинами. Если известно, что из этих 100 человек 10 увлекаются искусством, то в случае, если увлечение не зависит от пола, мы будем ожидать, что из 10 увлекающихся искусством 3 являются мужчинами, а 7 — женщинами. Сопоставляя эти ожидаемые частоты с наблюдаемыми частотами, мы можем судить о том, действительно ли два номинативных признака не связаны. Чем больше расхождение наблюдаемых и ожидаемых частот, тем, очевидно, два признака сильнее связаны друг с другом. Целью применения критерия независимости $x^2$ и является установление степени соответствия между наблюдаемыми и ожидаемыми значениями ячеек.

В основе критерия независимости лежит вычисление величины $\chi^2$, определяемой как сумма отношений суммы квадратов отклонений наблюдаемой величины /о от ожидаемой величины $f_e$ к ожидаемой величине каждой ячейки:

$$\chi^2 = \sum[(f_0 - f_e)^2 / f_e].$$

Как можно видеть из формулы, при больших отклонениях $f_0$ от $f_e$ величина $\chi^2$ также становится большой. Вместе с $\chi^2$ вычисляется $p$-уровень значимости. При $p > 0{,}05$ считается, что различия между наблюдаемыми и ожидаемыми значениями незначительны. В противном случае предположение о независимости двух номинативных переменных отклоняется и делается вывод о том, что две классификации (переменные) зависят друг от друга. Более подробные описания

величин, связанных с критерием $\chi^2$, вы можете найти в разделе «Представление результатов» этой главы.

Зачастую величина $\chi^2$ ошибочно воспринимается исследователями как величина *силы связи* между переменными. Однако это не так, поскольку $\chi^2$ в значительной степени определяется числом переменных таблицы сопряженности и размером выборки. Таким образом, сравнение двух значений $\chi^2$, полученных при разных условиях, становится бессмысленным. По этой причине Пирсон (Pearson) предложил коэффициент «фи», получаемый извлечением квадратного корня из отношения $x^2$ к размеру выборки *N*. Целью введения новой величины было получение наглядной интерпретации связи между переменными в виде коэффициента, лежащего в пределах от 0 до 1 и принимающего нулевое значение для независимых переменных и единичное значение для строго связанных переменных. Однако цель не была достигнута полностью: если одна из переменных таблицы сопряженности имеет более двух градаций, значение «фи» может превышать 1. Крамеру (Cramer) удалось исправить последний недостаток путем введения коэффициента $V = \sqrt{\chi^2 / [N(k-1)]}$, где $k$ — наименьшее из числа градаций двух переменных. Этот коэффициент всегда принимает значения от 0 до 1 и служит характеристикой силы связи между переменными.

# Пошаговый алгоритм вычислений

Для применения команды Crosstabs (Таблицы сопряженности) мы снова обратимся к файлу ex01.sav. Число объектов, или наборов значений, в этом файле равно 100, поэтому программа будет считать *N* равным 100. Мы создадим несколько таблиц сопряженности, а также применим критерий $\chi^2$ для определения зависимости переменных пол и хобби, но сначала необходимо выполнить три подготовительных шага.

**Шаг 1** Создайте новый файл данных или подготовьте существующий. Все необходимые действия описаны в главе 3.

**Шаг 2** Запустите программу SPSS при помощи значка на рабочем столе или команды Пуск ▸ Программы ▸ SPSS for Windows ▸ SPSS 11.5 for Windows (Start ▸ Programs ▸ SPSS for Windows ▸ SPSS 11.5 for Windows) главного меню Windows и в открывшемся после запуска программы диалоговом окне SPSS for Windows щелкните на кнопке Cancel (Отмена).

После выполнения этого шага на экране появится окно редактора данных SPSS.

**Шаг 3** Откройте файл данных, с которым вы намерены работать (в нашем случае это файл ex01.sav). Если он расположен в текущей папке, то выполните следующие действия:

1. Выберите в меню File (Файл) команду Open ▸ Data (Открытие ▸ Данные) или щелкните на кнопке Open File (Открытие файла) панели инструментов.
2. В открывшемся диалоговом окне дважды щелкните на имени ex01.sav или введите его с клавиатуры и щелкните на кнопке OK.

Независимо от того, открыта программа SPSS только что или какие-то процедуры уже выполнялись, в верхней части главного окна должна присутствовать строка меню (она показана ниже). Пока строка меню присутствует на экране, доступны все команды анализа данных. При этом окно с данными видеть не обязательно.

File Edit View Data Transform Analyze Graphs Utilities Window Help

При работе с таблицами результатов или при редактировании диаграмм некоторые пункты строки меню могут исчезать или меняться. Чтобы вернуться к главному окну и стандартной строке меню, щелкните на кнопке свертывания или восстановления текущего окна.

После завершения шага 3 на экране должно присутствовать окно редактора данных со строкой меню.

Шаг 4 ft В меню Analyze (Анализ) выберите команду Descriptive Statistics ▸ Crosstabs (Описательные статистики ▸ Таблицы сопряженности). На экране появится диалоговое окно Crosstabs (Таблицы сопряженности), показанное на рис. 8.1.

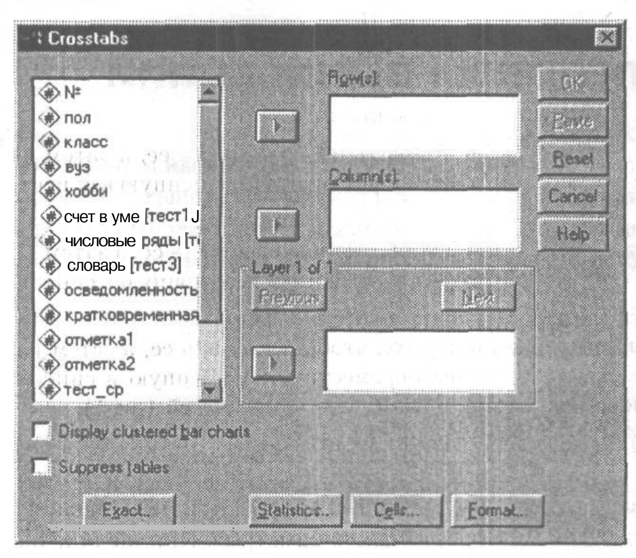



**Рис. 8.1.** Диалоговое окно Crosstabs

С помощью диалогового окна Crosstabs (Таблицы сопряженности) можно задать конфигурацию будущей таблицы сопряженности. Главное, что требуется сделать в этом окне, — это выбрать переменные в исходном списке и переместить их в целевые списки Row(s) (Строки) и Column(s) (Столбцы), используя верхнюю и среднюю кнопки со стрелкой. Тем самым вы определите, градации какой переменной будут составлять строки таблицы, а какой — столбцы. Поместив в список Row(s) (Строки) переменную пол, в список Column(s) (Столбцы) -

переменную хобби и щелкнув затем на кнопке OK, вы создадите таблицу с двумя строками и тремя столбцами.

Самый нижний список диалогового окна позволяет построить таблицу сопряженности для трех и более переменных. Если бы мы захотели подключить к таблице, содержащей переменные пол и хобби, еще и переменную класс, последнюю следовало бы поместить в нижний список при помощи нижней кнопки со стрелкой. В результате вы увидели бы на экране таблицу, разбитую на три части, каждая из которых представляла бы собой таблицу сопряженности пол×хобби для соответствующего значения переменной класс (1, 2 и 3). Кнопки Previous (Предыдущий) и Next (Следующий) слева и справа от метки Layer 1 of 1 (Слой 1 из 1) предназначены для случаев, когда необходимо построить таблицы сопряженности с более чем одной дополнительной переменной. Например, если бы вам понадобилось составить таблицу сопряженности с переменными пол, хобби, класс и вуз, то в нижнем списке нужно было бы указать две переменные класс и вуз, а в итоге программой была бы сгенерирована общая таблица сопряженности, включающая в себя 3 таблицы пол×хобби для каждой градации переменной класс и 4 таблицы пол×хобби для каждого значения переменной вуз.

В следующем примере мы создадим таблицу сопряженности пол×хобби×класс.

Шаг 5 При открытом диалоговом окне Crosstabs (Таблицы сопряженности), представленном на рис. 8.1, выполните следующие действия:

1. Щелкните сначала на переменной пол, чтобы выделить ее, а затем – на верхней кнопке со стрелкой, чтобы переместить переменную в список Row(s) (Строки).
2. Щелкните сначала на переменной хобби, чтобы выделить ее, а затем — на средней кнопке со стрелкой, чтобы переместить переменную в список Column(s) (Столбцы).
3. Щелкните сначала на переменной класс, чтобы выделить ее, а затем на нижней кнопке со стрелкой, чтобы переместить переменную в список дополнительных переменных под меткой Layer 1 of 1 (Слой 1 из 1).
4. Щелкните на кнопке OK, чтобы открыть окно вывода.

Как правило, исследователи редко ограничиваются вычислением частот, предпочитая дополнять процедуру заданием разнообразных параметров. При щелчке на кнопке Cells (Ячейки) открывается диалоговое окно Crosstabs: Cell Display (Таблицы сопряженности: Ячейки), представленное на рис. 8.2. Это окно позволяет управлять информацией в ячейках. По умолчанию флажок Observed (Наблюдаемые) в группе Counts (Значения) установлен, поскольку наблюдаемые частоты являются главной вычисляемой величиной в любой таблице сопряженности. Установка флажка Expected (Ожидаемые) позволит вычислять в ячейках и ожидаемые частоты. Остальные параметры задаются в зависимости от целей и предпочтений конкретного исследователя.

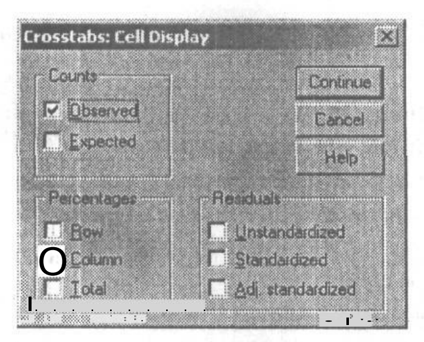



**Рис. 8.2.** Диалоговое окно Crosstabs: Cell Display

Ниже описано назначение некоторых флажков, управляющих содержимым ячеек в диалоговом окне Crosstabs: Cell Display (Таблицы сопряженности: Ячейки).

- При установке флажка Observed (Наблюдаемые) в группе Counts (Значения) отображается реальное количество объектов, соответствующее каждой ячейке.
- При установке флажка Expected (Ожидаемые) в группе Counts (Значения) отображается значение ожидаемой частоты для каждой ячейки, вычисленное в предположении, что соответствующие переменные являются независимыми.
- При установке флажка Row (Строки) в группе Percentages (Проценты) отображается частота ячейки в процентах от суммарной частоты строки, в которой она находится.
- При установке флажка Column (Столбцы) в группе Percentages (Проценты) отображается частота ячейки в процентах от суммарной частоты столбца, в котором она находится.
- При установке флажка Total (Всего) в группе Percentages (Проценты) отображается частота ячейки в процентах от суммарной частоты всей таблицы сопряженности.
- При установке флажка Unstandardized (Нестандартизированные) в группе Residuals (Остатки) отображается разность между наблюдаемой и ожидаемой частотами.

В следующем примере мы создадим таблицу сопряженности полхxобби, в ячейки которой включим наблюдаемые и ожидаемые частоты, процент от суммы и нестандартизированный остаток.

Обратите внимание, что мы не устанавливали флажок Observed (Наблюдаемые), поскольку он уже установлен программой по умолчанию.

В приведенных примерах мы научились создавать таблицы сопряженности и заполнять их ячейки различными показателями. Как правило, на практике при создании таблиц сопряженности применяется критерий $\chi^2$. Для настройки параметров анализа необходимо в диалоговом окне Crosstabs (Таблицы сопряженности) щелкнуть на кнопке Statistics (Статистики). В ответ программа откроет диалоговое окно

Crosstabs: Statistics (Таблицы сопряженности: Статистики), представленное па рис. 8.3. В этом окне с помощью флажков можно выбрать желаемые статистические критерии и меры независимости и согласия распределений. В следующих примерах мы расскажем лишь о трех наиболее употребительных флажках: Chi-square (Хи-квадрат), Phi and Cramer's V (φ и V Крамера) и Correlations (Корреляции). Как и в случае окна Cells (Ячейки), вам необходимо установить нужные флажки, а затем вернуться в диалоговое окно Crosstabs (Таблицы сопряженности) для завершения анализа. Описания остальных критериев и показателей, флажки для которых есть в диалоговом окне Crosstabs: Statistics (Таблицы сопряженности: Статистика), можно найти в справочной системе или документации к программе SPSS for Windows.

**Шаг 5а** После выполнения шага 4 у вас должно быть открыто диалоговое окно Crosstabs (Таблицы сопряженности), представленное на рис. 8.1. Если вы уже успели поработать с этим окном, очистите его щелчком на кнопке Reset (Сброс) и выполните следующие действия:

1. Щелкните сначала па переменной пол, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в список Row(s) (Строки).
2. Щелкните сначала на переменной хобби, чтобы выделить ее, а затем — на средней кнопке со стрелкой, чтобы переместить переменную в список Column(s) (Столбцы).
3. Щелкните на кнопке Cells (Ячейки), чтобы открыть диалоговое окно Crosstabs: Cell Display (Таблицы сопряженности: Ячейки), и установите флажки Expected (Ожидаемые), Total (Всего) и Unstandardized (Нестандартизированные).
4. Щелкните па кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Crosstabs (Таблицы сопряженности).
5. Щелкните на кнопке OK, чтобы открыть окно вывода.

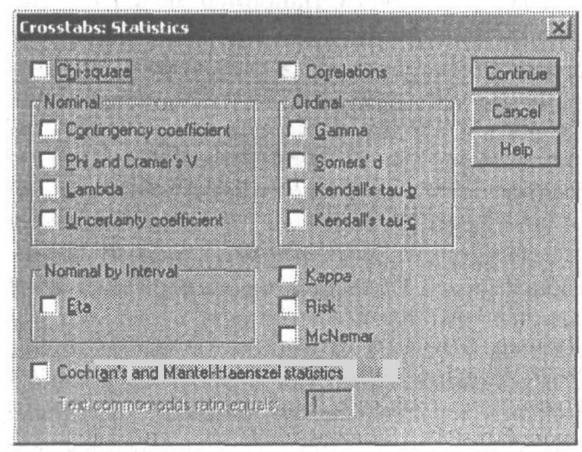



**Рис. 8.3.** Диалоговое окно Crosstabs: Statistics

В следующем примере мы создадим таблицу сопряженности полххобби, в ячейки которой включим наблюдаемые и ожидаемые частоты, нестандартизированные остатки, применим критерий $\chi^2$ и вычислим коэффициенты ер и V Крамера. Анализ корреляций между переменными в данном случае не имеет смысла, так как они не являются количественными.

**Шаг 5б** После выполнения шага 4 у вас должно быть открыто диалоговое окно Crosstabs (Таблицы сопряженности), представленное па рис. 8.1. Если вы уже успели поработать с этим окном, очистите его щелчкомна кнопке Reset (Сброс) и выполните следующие действия:

1. Щелкните сначала на переменной пол, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в список Row(s) (Строки).
2. Щелкните сначала на переменной хобби, чтобы выделить ее, а затем — на средней кнопке со стрелкой, чтобы переместить переменную в список Column(s) (Столбцы).
3. Щелкните на кнопке Cells (Ячейки), чтобы открыть диалоговое окно Crosstabs: Cell Display (Таблицы сопряженности: Ячейки), и установите флажки Expected (Ожидаемые) и Unstandardized (Нестандартизированные).
4. Щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Crosstabs (Таблицы сопряженности).
5. Щелкните на кнопке Statistics (Статистики), чтобы открыть диалоговое окно Crosstabs: Statistics (Таблицы сопряженности: Статистики), и установите флажки Chi-square (Хи-квадрат) и Phi and Cramer's V (Фи и V Крамера).
6. Щелкните па кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Crosstabs (Таблицы сопряженности).
7. Щелкните на кнопке OK, чтобы открыть окно вывода.

Зачастую приходится строить таблицы сопряженности и применять критерий $x^2$ не для всей выборки объектов, а только для их части. Например, может возникнуть необходимость в создании таблицы полххобби только для двух из трех градаций переменной хобби: спорт и искусство. Это делается следующим образом. После выбора переменных для построения перекрестной таблицы, определения набора величии, вычисляемых для каждой ячейки, и задания желаемых статистик, в меню Data (Данные) выберите команду Select Cases (Выбор объектов). На экране появится одноименное диалоговое окно (см. раздел «Выбор объектов для анализа» в главе 4), в котором в группе Select (Выбор) следует установить переключатель If condition is satisfied (Если удовлетворяет условию), а затем щелкнуть на расположенной рядом кнопке If (Если). Откроется диалоговое окно Select Cases: If (Выбор объектов: Если). Элементы интерфейса этого окна позволяют выполнять операции, о которых шла речь в главе 4. В данном случае пас будет интересовать выбор уровней 1 и 3 переменной хобби. Для исключения уровня 2 из анализа выделите переменную хобби в исходном списке, введите ее в условие отбора, щелкнув на кнопке с направленной вправо стрелкой, на панели калькулятора щелкните на кнопке ~= (не равно) и введите в условие отбора число 2.

После этого щелкните сначала на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Select Cases (Выбор объектов), а затем — на кнопке ОК, чтобы вернуться в диалоговое окно Crosstabs (Таблицы сопряженности). Наконец, щелчок на кнопке ОК в окне Crosstabs (Таблицы сопряженности) приведет к построению таблицы сопряженности с двумя выбранными градациями переменной хобби и двумя градациями переменной пол.

Реализуем описанный алгоритм в виде пошаговых инструкций.

**Шаг 5в** После выполнения шага 4 у вас должно быть открыто диалоговое окно Crosstabs (Таблицы сопряженности), представленное на рис. 8.1. Если вы уже успели поработать с этим окном, очистите его щелчком на кнопке Reset (Сброс) и выполните следующие действия:

1. Щелкните сначала на переменной пол, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в список Row(s) (Строки).
2. Щелкните на переменной хобби, чтобы выделить ее, а затем — на средней кнопке со стрелкой, чтобы переместить переменную в список Column(s) (Столбцы).
3. Щелкните на кнопке Cells (Ячейки), чтобы открыть диалоговое окно Crosstabs: Cell Display (Таблицы сопряженности: Ячейки), и установите флажки Expected (Ожидаемые) и Unstandardized (Нестандартизированные).
4. Щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Crosstabs (Таблицы сопряженности).
5. Щелкните на кнопке Statistics (Статистики), чтобы открыть диалоговое окно Crosstabs: Statistics (Таблицы сопряженности: Статистики), и установите флажки Chi-square (Хи-квадрат) и Phi and Cramer's V (φ и V Крамера).
6. Щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Crosstabs (Таблицы сопряженности).
7. В меню Data (Данные) выберите команду Select Cases (Выбор объектов), чтобы открыть одноименное диалоговое окно, установите переключатель If condition is satisfied (Если удовлетворяет условию) и щелкните на расположенной рядом кнопке If (Если), открыв диалоговое окно Select Cases: If (Выбор объектов: Если).
8. В списке доступных переменных щелкните сначала на переменной хобби, чтобы выделить ее, а затем — на кнопке с направленной вправо стрелкой, чтобы ввести переменную в условие отбора. Дополните условие отбора оператором ~= (не равно) и числом 2, используя панель калькулятора и клавиатуру, и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Select Cases (Выбор объектов).
9. Щелкните на кнопке ОК, чтобы вернуться в диалоговое окно Crosstabs (Таблицы сопряженности).
10. Щелкните на кнопке ОК, чтобы открыть окно вывода.

После выполнения шага 5, 5а, 5б или 5в программа автоматически активизирует окно вывода. Для просмотра результатов вы при необходимости можете

воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

# Печать результатов и выход из программы

Ниже описана типичная процедура печати результатов статистического анализа (или нескольких анализов). После выполнения шага 5 должно быть открыто окно вывода.

Шаг 6 В окне вывода укажите фрагменты, выводимые на печать (см, раздел «Окно вывода» в главе 2), в меню File (Файл) выберите команду Print (Печать), при необходимости задайте параметры печати и щелкните на кнопке OK.

Последнее, что необходимо сделать после завершения обработки и печати результатов, — это выйти из программы SPSS.

Шаг 7 Для выхода из программы в меню File (Файл) выберите команду Exit (Выход).

Иногда после выполнения команды Exit (Выход) на экране могут появляться небольшие диалоговые окна с вопросом о необходимости сохранения сделанных в файлах изменений и кнопками, описывающими возможные варианты ответа. Для завершения работы просто щелкайте на соответствующих кнопках.

# Представление результатов

На рис. 8.4 приведены фрагменты выводимых результатов, сгенерированные программой после выполнения шага 5б.

Первое, на что необходимо обратить внимание при интерпретации выводимых результатов, — это соотношение между наблюдаемыми (Count) и ожидаемыми (Expected count) частотами в каждой ячейке. Беглый взгляд на первую таблицу говорит о том, что эти величины заметно различаются. Как уже отмечалось, это различие характеризуется остатком, который рассчитывается как разность между наблюдаемой и ожидаемой величинами. При составлении таблицы сопряженности мы включили в каждую ячейку остаток (Residual), поэтому на самом деле вам нет необходимости самому сравнивать наблюдаемую величину с ожидаемой, чтобы оценить их соотношение. Итак, различие наблюдаемых и ожидаемых величин говорит о том, что переменные пол и хобби связаны друг с другом. Это подтверждается относительно большим значением критерия $x^2$ (14,032) и малым

значением $p$-уровня ($p$ — 0,000), то есть высокой статистической значимостью. Как можно видеть, аналогичная картина наблюдается и с результатами остальных тестов па независимость.

**ПОЛ * ХОББИ Crosstabulation**

| | | | ХОББИ | | Total |
|---|---|---|---|---|---|
| | | | спорт | искусство | |
| ПОЛ | ЖЕН | Count | 15 | 27 | 42 |
| | | Expected Count | 22.0 | 20.0 | 42.0 |
| | | Residual | -7.0 | 7.0 | |
| | МУЖ | Count | 18 | 3 | 21 |
| | | Expected Count | 11.0 | 10.0 | 21.0 |
| | | Residual | 7.0 | -7.0 | |
| Total | | Count | 33 | 30 | 63 |
| | | Expected Count | 33.0 | 30.0 | 63.0 |

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
|---|---|---|---|---|---|
| Pearson Chi-Square | 14.032[b] | 1 | .000 | | |
| Continuity Correction[a] | 12.099 | 1 | .001 | | |
| Likelihood Ratio | 15.221 | 1 | .000 | | |
| Fisher's Exact Test | | | | .000 | .000 |
| Linear-by-Linear Association | 13.809 | 1 | .000 | | |
| N of Valid Cases | 63 | | | | |

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 10.00.

**Symmetric Measures**

| | | Value | Approx. Sig. |
|---|---|---|---|
| Nominal by Nominal | Phi | -.472 | .000 |
| | Cramer's V | .472 | .000 |
| N of Valid Cases | | 63 | |

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

**Рис. 8.4.** Фрагменты окна вывода после выполнения шага 5б

Для таблиц 2 × 2 при небольшом числе объектов (до 40-50) следует применять критерий $\chi^2$ с поправкой на непрерывность (Continuity Correction) Йетса. Поэтому более корректным в нашем случае будет выбрать именно такой вариант критерия.

Для таблиц сопряженности большой размерности следует помнить о проблеме малых значений частот. Малым считается значение ожидаемой частоты, меньшее 5. Предполагается, что если более 25 % ячеек таблицы сопряженности имеют малые ожидаемые значения частот, то вероятность ошибки $\chi^2$-анализа очень высока. В этом случае необходимо исключать градации с малыми значениями частот или объединять градации.

# Терминология, используемая при выводе

Трактовка терминов, используемых программой в окне вывода, дана далее.

- Count (Наблюдаемая частота) — верхнее из чисел, приведенных в каждой ячейке, равное количеству объектов в соответствующей градации.
- Expected Count (Ожидаемая частота) — среднее из чисел, приведенных в каждой ячейке, равное количеству объектов в соответствующей градации, вычисленное в предположении, что переменные, входящие в таблицу сопряженности, являются независимыми.
- Residual (Остаток) — разность между наблюдаемой и ожидаемой частотами.
- Total (Всего) — суммарное число объектов в каждой строке (42 — для женщин, 21 — для мужчин) или в каждом столбце (33 — спорт, 30 — искусство).
- Pearson Chi-Square (Хи-квадрат Пирсона) и Continuity Correction (С поправкой на непрерывность) — два варианта критерия $\chi^2$. При больших значениях $N$ эти коэффициенты практически равны. Формула для $x^2$ Пирсона выглядит следующим образом:

$$\chi^2 = \sum[(f_0 - f_e)^2 / f_e].$$

- Value (Значение) — для критерия $\chi^2$ значение тем больше, чем больше зависимость между переменными (как в нашем примере). Значения, близкие к О, свидетельствуют о независимости переменных.
- df (Степени свободы) — произведение количества градаций переменных, уменьшенных на 1. Для дайной таблицы сопряженности число степеней свободы равно

$$(2 - 1) \times (2 - 1) = 1.$$

- Asymp. Sig. (Асимптотическая значимость) — мера случайности связи, или $p$-уровень значимости, то есть вероятность того, что связь является случайной. Чем меньше эта величина, тем выше статистическая значимость (достоверность) связи.
- Linear-by-linear Association (Линейная связь между переменными) — статистический критерий, определяющий степень корреляции между переменными. Чаще всего результаты этого теста являются бессмысленными, поскольку порядок уровней переменных не имеет логической интерпретации. К примеру, нет смысла говорить об упорядоченной структуре хобби. Тем не менее, если бы одной из переменных был уровень дохода, упорядоченный по возрастанию, количественная оценка корреляции несла бы вполне определенную смысловую нагрузку.

- Minimum expected count (Минимальная ожидаемая частота) — смысл этого значения в полной мере раскрывается его названием. В рассматриваемой таблице сопряженности минимальная ожидаемая частота равна 10 — для юношей, увлекающихся искусством.

- Nominal by Phi (Значение ф) — коэффициент, являющийся мерой связи двух переменных и вычисляющийся по формуле $\varphi = (\chi^2/N)^{1/2}$. Значение ф = -0,472 показывает умеренную связь между двумя переменными.

- Nominal Cramer's V (Значение V Крамера) — как и коэффициент ф, этот коэффициент является мерой связи между двумя переменными, однако отличается тем, что всегда принимает значения от 0 до 1. Коэффициент Крамера определяется следующей формулой:

$$V = \sqrt{\chi^2 \ / \ [N(k-1)]}.$$

- Approx. sig. (Приблизительная значимость) — как и для критерия $\chi^2$, является мерой значимости вычисленного значения. Значимость, равная 0,000 (как в нашем примере), говорит о высокой статистической значимости связи.

# 9 Корреляции



Для вычисления корреляций между данными в программе SPSS используются команды подменю Correlate (Корреляция) меню Analyze (Анализ). Корреляция представляет собой величину, заключенную в пределах от -1 до +1, и обозначается буквой *r*. Понятия *корреляция* и *двумерная корреляция* часто употребляются как синонимы; последнее означает «корреляция между двумя переменными» и подчеркивает, что рассматривается именно двухмерное соотношение. Основной коэффициент корреляции *r* Пирсона предназначен для оценки связи между двумя переменными, измеренными в метрической шкале, распределение которых соответствует нормальному. Несмотря на то что величина *r* рассчитывается в предположении, что значения обеих переменных распределены по нормальному закону, формула для ее вычисления дает достаточно точные результаты и в случаях аномальных распределений, а также в случаях, когда одна из переменных является дискретной. Для распределений, не являющихся нормальными, предпочтительнее пользоваться ранговыми коэффициентами корреляции Спирмена или Кендалла. Команды подменю Correlate (Корреляция) позволяют вычислить как коэффициент Пирсона (Pearson), так и коэффициенты Спирмена (Spearman) и Кендалла (Kendall's tau-b). Существуют и другие коэффициенты корреляции, применяющиеся для самых разных типов данных, однако их описание выходит за рамки темы этой книги.

## Понятие корреляции

*Корреляция,* или *коэффициент корреляции,* — это статистический показатель вероятностной связи между двумя переменными, измеренными в количественной шкале. В отличие от функциональной связи, при которой каждому значению одной переменной соответствует строго определенное значение другой переменной, вероятностная связь характеризуется тем, что каждому значению одной переменной соответствует множество значений другой переменной. Примером вероятностной связи является связь между ростом и весом людей. Ясно, что один и тот же рост может быть у людей разного веса, как и наоборот. Величина коэффици-

ента корреляции меняется от -1 до 1. Крайние значения соответствуют линейной функциональной связи между двумя переменными, 0 — отсутствию связи.

Наглядное представление о связи двух переменных дает график двухмерного рассеивания — соответствующая команда Scatter (Рассеивание) имеется в меню Graphs (Графики). На таком графике каждый объект представляет собой точку, координаты которой заданы значениями двух переменных. Таким образом, множество объектов представляет собой на графике множество точек. По конфигурации этого множества точек можно судить о характере связи между двумя переменными.

*Строгая положительная корреляция* (perfect positive correlation) определяется значением $r$ - 1. Термин «строгая» означает, что значения одной переменной однозначно определяются значениями другой переменной, а термин «положительная» — что с возрастанием значений одной переменной значения другой переменной также возрастают.

Строгая корреляция является математической абстракцией и практически не встречается в реальных исследованиях. Примером строгой корреляции является соответствие между временем пути и пройденным расстоянием при неизменной скорости.

*Положительная корреляция* соответствует значениям $0 < r < 1$. Положительную корреляцию следует интерпретировать следующим образом: если значения одной переменной возрастают, то значения другой имеют *тенденцию* к возрастанию. Чем коэффициент корреляции ближе к 1, тем сильнее эта тенденция, и обратно, с приближением коэффициента корреляции к 0 тенденция ослабевает.

Примером значительной положительной корреляции служит зависимость между ростом и весом человека. Считается, что в этом случае коэффициент корреляции равен $r = 0{,}83$. Слабая положительная корреляция ($r = 0{,}12$) наблюдается между способностью человека к сочувствию и реальной помощью, которую он оказывает нуждающимся людям.

*Отсутствие корреляции* (no correlation) определяется значением $r = 0$. Нулевой коэффициент корреляции говорит о том, что значения переменных никак не связаны друг с другом. Примером пары величин с нулевой корреляцией является рост человека и результат его IQ-теста.

*Отрицательная корреляция* соответствует значениям $-1 < r < 0$. Если значения одной переменной возрастают, то значения другой имеют *тенденцию* к убыванию. Чем коэффициент корреляции ближе к -1, тем сильнее эта тенденция, и обратно, с приближением коэффициента корреляции к 0 тенденция ослабевает.

Слабая отрицательная корреляция ($r = -0{,}13$) наблюдается между агрессивностью человека по отношению к своему другу и помощью, которую он ему оказывает. Чем агрессивней человек, тем помощь меньше, однако зависимость выражена слабо. Примером значительной отрицательной корреляции ($r = -0{,}73$) служит зависимость между нервной возбудимостью человека и его эмоциональной уравновешенностью. Чем выше оказывается результат его теста па возбудимость, тем более низкий результат имеет его тест на уравновешенность.

*Строгая отрицательная корреляция* (perfect negative correlation) определяется значением $r = -1$. Она, так же как и строгая положительная корреляция, является абстракцией и не находит отражения в практических исследованиях. Пример, иллюстрирующий строгую отрицательную корреляцию, можно взять из школьного учебника физики: при равномерном движении расстояние равно произведению времени на скорость. При заданном расстоянии время и скорость являются обратно пропорциональными величинами: чтобы пройти путь за половину времени, необходимо идти вдвое быстрее.

# Дополнительные сведения

## Линейная и криволинейная корреляции

Основной коэффициент корреляции $r$ Пирсона является мерой прямолинейной связи между переменными: его значения достигают максимума, когда точки на графике двухмерного рассеивания лежат на одной прямой линии. В реальной жизни отношения между переменными часто оказываются не только вероятностными, но и непрямолинейными: монотонными или немонотонными. Если связь нелинейная, но монотонная, то вместо $r$ Пирсона следует использовать ранговые корреляции Спирмена или Кендалла.

Нередко связь между двумя переменными является не только нелинейной, но и немонотонной. В качестве примера рассмотрим такие два фактора, как нервное возбуждение перед экзаменом и успешность его сдачи. Исследования показывают, что студенты, испытывающие умеренное нервное возбуждение, имеют наилучшие результаты на экзаменах, в то время как очень спокойные или очень нервные студенты сдают экзамены значительно хуже. Если по оси абсцисс отложить степень нервного возбуждения, а по оси ординат — результаты сдачи экзаменов, то график зависимости между ними примет вид, близкий к перевернутой букве U. При этом любой коэффициент корреляции, вычисленный для этих величин, окажется весьма низким. Это объясняется тем, что для немонотонных отношений нужны другие методы оценки корреляции. Частично мы коснемся этих методов в главах 15 и 16, посвященных видам регрессионного анализа.

Перед тем как оценивать корреляцию двух переменных, рекомендуется построить график зависимости между ними — график двухмерного рассеивания. Если график демонстрирует монотонность связи, то для вычисления корреляции можно использовать команды подменю Correlate (Корреляция).

## Ранговые корреляции

Как уже отмечалось, необходимость в применении ранговых корреляций возникает в двух случаях: когда распределение хотя бы одной из двух переменных не соответствует нормальному и когда связь между переменными является нелинейной (но монотонной). В этих случаях вместо корреляции $r$ Пирсона можно

выбрать ранговые корреляции: г Спирмена либо т (читается «тау») Кендалла. Ранговыми они являются потому, что программа предварительно ранжирует переменные, между которыми они вычисляются.

Корреляцию *r* Спирмена программа SPSS вычисляет следующим образом: сначала переменные переводятся в ранги, а затем к рангам применяется формула *r* Пирсона. Таким образом, *r* Спирмеиа интерпретируется по аналогии с *r* Пирсона. Иначе дело обстоит с корреляцией т Кендалла, которая имеет вероятностную природу.

Рассмотрим принцип вычисления т Кендалла на примере. Предположим, оценивается связь между ростом и весом в группе людей, предварительно ранжированных по этим переменным. Тогда при сравнении любых двух человек из этой группы возможны две ситуации: однонаправленное изменение переменных («совпадение»), когда и рост, и вес одного больше, чем другого, и разнонаправленное изменение («инверсия»), когда рост у второго больше, а вес меньше, чем у первого. Перебрав все пары испытуемых, можно оценить вероятность совпадений (P) и вероятность инверсий (*Q*). Корреляция Кендалла — это разность вероятностей «совпадений» и «инверсий»: т = *P* - *Q*. По значению корреляции Кендалла можно всегда вычислить вероятность «совпадений» (P = (1 + т)/2) и «инверсий» (*Q* = (1 - τ)/2). Например, если корреляция между ростом и весом т - 0,5, то вероятность «совпадений» (чем больше рост, тем больше вес) P - 0,75, а вероятность «инверсий» (чем больше рост, тем меньше вес) *Q* = 0,25. Таким образом, важным преимуществом корреляции т Кендалла является ее отчетливая вероятностная интерпретация.

## Значимость

Как и большинство статистических процедур, команды подменю Correlate (Корреляция) наряду с описательными статистиками (корреляциями в данном случае) вычисляют их уровень значимости. Напомним, что уровень значимости является мерой статистической достоверности результата вычислений, в данном случае — корреляции, и служит основанием для интерпретации. Если исследование показало, что уровень значимости корреляции не превышает 0,05, то это означает, что с вероятностью 5 % и менее корреляция является случайной. Обычно это является основанием для вывода о статистической достоверности корреляции. В противном случае ($p > 0{,}05$) связь признается статистически недостоверной и не подлежит содержательной интерпретации.

SPSS позволяет определять два теста значимости: односторонний (one-tailed) и двусторонний (two-tailed). Обычно используется двусторонний тест значимости. Но если вы заранее знаете направление корреляции (положительное или отрицательное) и вас интересует только одно направление, то можно использовать односторонний тест значимости. Однако такая ситуация встречается редко, а если и встречается, то правомерность односторонней проверки с трудом поддается обоснованию.

## Частная корреляция

Понятие *частной корреляции* (partial correlation) связано с ковариацией, разговор о которой пойдет в главе 14. Здесь мы упоминаем частную корреляцию лишь как одну из команд подменю Correlate (Корреляция). Суть частной корреляции заключается в следующем. Если две переменные коррелируют, то всегда можно предположить, что эта корреляция обусловлена влиянием третьей переменной, как общей причины совместной изменчивости первых двух переменных. Для проверки этого предположения достаточно исключить влияние этой третьей переменной и вычислить корреляцию двух переменных без учета влияния третьей переменой (при фиксированных ее значениях). Корреляция, вычисленная таким образом, и называется частной. Например, при исследовании связи между скоростью чтения и зрелостью моральных суждений у детей разного возраста наверняка будет обнаружена корреляция этих двух переменных. Ответ на вопрос, связаны ли они непосредственно, или связь обусловлена возрастом, позволяет дать частная корреляция. Если при фиксированных значениях возраста частная корреляция скорости чтения и зрелости моральных суждений приближается к нулю, то можно заключить, что связь между этими переменными обусловлена возрастом.

# Пошаговые алгоритмы вычислений

В пошаговых процедурах приведено несколько примеров, иллюстрирующих применение команд подменю Correlate (Корреляция). В качестве файла данных используется файл ex01.sav с входящими в него пятью переменными тест1, ..., тест5. Сначала выполняются три подготовительных шага.

**Шаг 1** Создайте новый файл данных или подготовьте существующий. Все необходимые действия описаны в главе 3.

**Шаг 2** Запустите программу SPSS при помощи значка на рабочем столе или команды Пуск ▶ Программы ▶ SPSS for Windows ▶ SPSS 11.0 for Windows (Start ▶ Programs ▶ SPSS for Windows ▶ SPSS 11.0 for Windows) главного меню Windows. В открывшемся после запуска программы диалоговом окне SPSS for Windows щелкните на кнопке Cancel (Отмена).

После выполнения этого шага па экране появится окно редактора данных SPSS.

**Шаг 3** Откройте файл данных, с которым вы намерены работать (в нашем случае это файл ex01.sav). Если он расположен в текущей папке, то выполните следующие действия:

1. Выберите в меню File (Файл) команду Open ▶ Data (Открытие ▶ Данные) или щелкните па кнопке Open File (Открытие файла) панели инструментов.
2. В открывшемся диалоговом окне дважды щелкните на имени ex01.sav или введите его с клавиатуры и щелкните на кнопке OK.

Независимо от того, открыта программа SPSS только что или какие-то процедуры уже выполнялись, в верхней части главного окна должна присутствовать строка меню (она показана ниже). Пока строка меню присутствует на экране, доступны все команды анализа данных. При этом окно с данными видеть не обязательно.




При работе со сводными таблицами или при редактировании диаграмм некоторые пункты строки меню могут исчезать или меняться. Чтобы вернуться к главному окну и стандартной строке меню, щелкните на кнопке свертывания или восстановления текущего окна.

После завершения шага 3 на экране должно присутствовать окно редактора данных со строкой меню.

**Шаг 4** В меню Analyze (Анализ) выберите команду Correlate ▸ Bivariate (Корреляция ▸ Двумерная). На экране появится диалоговое окно Bivariate Correlations (Двумерные корреляции), показанное на рис. 9.1.

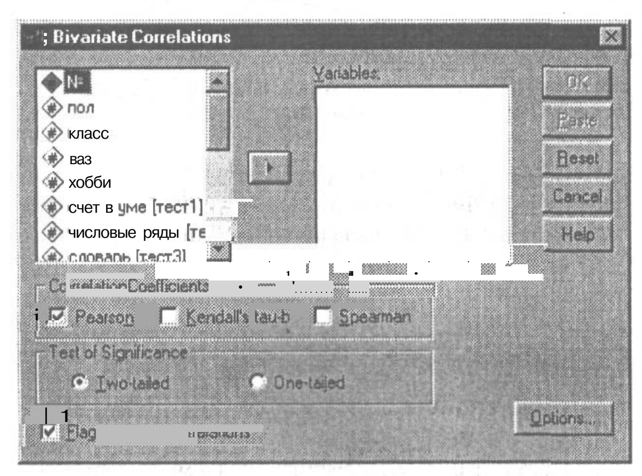



**Рис. 9.1.** Диалоговое окно Bivariate Correlations

Окно Bivariate Correlations (Двумерные корреляции) позволяет настраивать параметры вычисления корреляций. В списке слева содержатся имена всех переменных, имеющих числовой тип. Строчные переменные для команды Correlate ▸ Bivariate (Корреляция ▸ Двумерная) недоступны.

Чтобы выбрать переменные для вычисления корреляции, их требуется переместить в список Variables (Переменные) при помощи кнопки со стрелкой. Если несколько нужных переменных располагается в исходном списке друг за другом, вы можете, наведя указатель мыши на верхнюю из них и нажав кнопку мыши, переместить указатель на нижнюю переменную и отпустить кнопку мыши, тем самым выделив сразу несколько переменных.

*В* группе Correlation Coefficients (Коэффициенты корреляции) по умолчанию установлен флажок Pearson (Пирсон). Если требуется вычислить ранговые корреляции, то следует установить флажок Spearman (Спирмен) и (или) Kendall's tau-b (Тау-би Кендалла). Можете установить все три флажка, чтобы иметь возможность сравнивать три коэффициента корреляции для различных распределений данных.

В группе Test of Significance (Тест значимости) по умолчанию установлен переключатель Two-tailed (Двусторонний). Если вы заранее уверены в направлении (знаке) корреляции, то можете установить переключатель One-tailed (Односторонний).

Флажок Flag significant correlations (Помечать значимые корреляции) по умолчанию установлен. Это означает, что корреляции, вычисленные с уровнем значимости от 0,01 до 0,05, будут помечены одной звездочкой (*), а от 0 до 0,01 — двумя звездочками (**). Вне зависимости от значимости в вывод включаются коэффициенты корреляции и $p$-уровни, вычисленные с точностью до 3 знаков после запятой, а также количество объектов, участвовавших в процедуре.

В примерах, приведенных в этой главе, мы будем пользоваться коэффициентом корреляции Пирсона, двусторонним тестом значимости, а также помечать звездочками значимые корреляции. Если в ваших исследованиях понадобится использовать иную конфигурацию параметров, вы сами можете легко их настроить, устанавливая нужные флажки и переключатели. «Отправной точкой» при выполнении всех примеров служит диалоговое окно Bivariate Correlations (Двумерные корреляции).

**Шаг 5** На этом шаге мы создадим корреляционную матрицу для значений переменных тест1, ..., тест5. После выполнения шага 4 должно быть открыто диалоговое окно Bivariate Correlations (Двумерные корреляции), представленное иа рис 9.1.

1. Дважды щелкните иа переменной тест1, чтобы переместить ее в список Variables (Переменные).
2. Повторите предыдущее действие для переменных тест2, ..., тест5.
3. Щелкните на кнопке OK, чтобы открыть окно вывода.

Кнопка Options (Параметры) позволяет задать дополнительные параметры корреляции. При щелчке на этой кнопке открывается диалоговое окно Bivariate Correlations: Options (Двухмерные корреляции: Параметры), представленное па рис. 9.2.

В группе Statistics (Статистики) имеется два флажка, управляющих отображением статистических величии: Means and standard deviations (Средние значения и стандартные отклонения) и Cross-product deviations and covariances (Произведения отклонений и ковариации). Группа Missing Values (Пропущенные значения) из двух переключателей позволяет выбрать способ исключения объектов, содержащих пропущенные значения. Установка переключателя Exclude cases pairwise (Попарное исключение объектов) означает, что если при вычислении корреляции между парой переменных для какого-нибудь объекта обнаружится отсутствующее значение, то объект будет исключен из вычисления, но только для этой пары переменных. В результате может оказаться, что для разных пар переменных коэффициенты корреля-

ции будут вычислены с разным числом объектов. При установке переключателя Exclude cases listwise (Построчное исключение объектов) программа перед началом вычислительного процесса исключит из рассмотрения все объекты, содержащие хотя бы одно отсутствующее значение. В любом случае, разрешение проблемы отсутствующих значений лучше провести до начала анализа. О том, каким образом это можно сделать, рассказывается в главе 4.

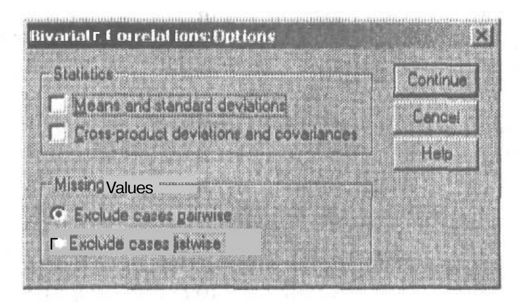



**Рис. 9.2.** Диалоговое окно Bivariate Correlations: Options

В приведенном примере программа генерирует квадратную корреляционную матрицу. Нередко исследователю необходимо вычислить не все корреляции, а только их часть. Например, можно представить ситуацию, когда сначала создается корреляционная матрица размером 12 x 12, а затем возникает необходимость ввести в анализ две новые переменные и вычислить коэффициент корреляции между ними и 12 предшествующими переменными. Возможности комщщ подменю Correlate (Корреляция) не позволяют этого делать — придется создавать командный файл.

Для создания командного файла нужно открыть окно *редактора синтаксиса,* показанное на рис. 9.3, выбрав команду New ▸ Syntax (Новый ▸ Синтаксис) в меню File (Файл).

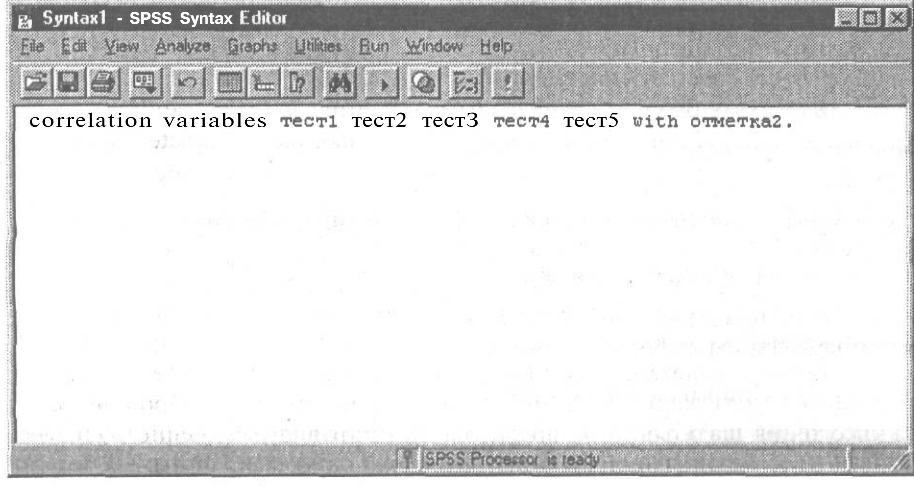



**Рис. 9.3.** Окно редактора синтаксиса SPSS

Необходимые команды вводятся непосредственно в окно редактора синтаксиса. Чтобы выполнить введенную команду, ее нужно выделить и щелкнуть на кнопке Run Current (Запуск команды) панели инструментов. Второй вариант запуска — выбрать в меню команду Run ▸ All (Запустить ▸ Все). Имейте в виду, что любой пропущенный знак, включая завершающую точку, или неверно написанное слово приведет к выдаче программой сообщения об ошибке.

**Шаг 5а** На этом шаге мы создадим корреляционную матрицу Пирсона размером 1 × 5, предназначенную для оценки степени зависимости переменной отметка2 от переменных тест1.....тест5.

1. В меню File (Файл) выберите команду New ▸ Syntax (Новый ▸ Синтаксис). В открывшееся окно редактора синтаксиса введите следующую команду (см. рис. 9.3):

```
correlations variables тест1 тест2 тест3 тест4 тест5 with отметка2.
```

2. Выберите команду Run ▸ All (Запустить ▸ Все) или выделите всю команду левой кнопкой мыши и щелкните на кнопке Run Current (Запуск команды) панели инструментов. Команда будет выполнена и откроется окно вывода.

В конструкции, приведенной здесь, можно произвольно изменять имена переменных. Допускается любое количество переменных как до, так и после ключевого слова with. Переменные, перечисленные до слова with, составят строки новой корреляционной матрицы, а после него — ее столбцы. Если вы получите на экране сообщение об ошибке, то, скорее всего, вы либо неверно ввели имя переменной, либо забыли указать точку в конце команды.

Ниже приведены варианты синтаксиса, позволяющие получать различные результаты обработки.

```
correlations variables тест1 тест2 тест3  тест4 тест5
  with отметка2 /missing listwise.
```

Вычисление корреляции Пирсона с учетом пропусков путем построчного удаления.

```
nonpar corr тест1 тест2 тест3 тест4 тест5 with отметка2.
```

Вычисление корреляции *r* Спирмена (с попарным удалением пропущенных значений).

```
nonpar corr тест1 тест2 тест3 тест4 тест5 with отметка2 /print kendall.
```

Вычисление корреляций τ Кендалла.

```
nonpar corr тест1 тест2 тест3 тест4 тест5
  with отметка2/missing listwise /print both.
```

Вычисление *r* Спирмена и τ Кендалла с построчным удалением пропусков.

После выполнения шага 5 или 5а программа автоматически активизирует окно вывода. Для просмотра результатов вы при необходимости можете воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание

на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

## Печать результатов и выход из программы

Ниже описана типичная процедура печати результатов статистического анализа (или нескольких анализов). После выполнения шагов 5 или 5а будет открыто окно вывода.

**Шаг 6** В окне вывода укажите фрагменты, выводимые на печать (см. раздел «Окно вывода» в главе 2), в меню File (Файл) выберите команду Print (Печать), при необходимости задайте параметры печати и щелкните на кнопке OK.

Последнее, что необходимо сделать после завершения исследования и печати результатов, — это выйти из программы SPSS.

**Шаг 7** Для выхода из программы в меню File (Файл) выберите команду Exit (Выход).

Иногда после выполнения команды Exit (Выход) на экране могут появляться небольшие диалоговые окна с вопросом о необходимости сохранения сделанных в файлах изменений и кнопками, описывающими возможные варианты ответа. Для завершения работы просто щелкайте па соответствующих кнопках.

## Представление результатов

На рис. 9.4 приведен фрагмент выводимых данных, сгенерированных программой после выполнения шага 5.

Обратите внимание на структуру представленной на рисунке таблицы (см. рис. 9.4). Верхнее значение в каждой ячейке таблицы является коэффициентом корреляции между соответствующими двумя переменными, вычисленным с точностью до трех знаков после запятой. Следующее значение является уровнем значимости, с которым было произведено вычисление коэффициента корреляции. Нижнее значение представляет собой число объектов, участвовавших в вычислении коэффициента корреляции. При наличии отсутствующих значений эта величина в разных ячейках может быть разной. Надписи под таблицей поясняют смысл символов * и **, используемых в качестве меток для некоторых вычисленных коэффициентов корреляции, и содержат информацию об уровне значимости.

Как легко видеть, на главной диагонали таблицы лежат единичные значения коэффициентов корреляции для всех переменных. Смысл этого результата прост: каждая переменная имеет строгую положительную корреляцию сама с собой.

**Correlations**

| | | счет в уме | числовые ряды | сломрь | осведом ленность | кратковреме нная память |
|---|---|---|---|---|---|---|
| счет • уме | Pearson Correlation | 1 | .436" | -.051 | -.134 | ,017 |
| | Slg. (2-talled) | . | ,000 | .617 | .166 | .867 |
| | N | 100 | 100 | 100 | 100 | 100 |
| числовые ряды | Pearson Correlation | .436" | 1 | -.196 | -.153 | .016 |
| | Slg. (2-talled) | .000 | . | .050 | .127 | .886 |
| | N | 100 | 100 | 100 | 100 | 100 |
| словарь | Pearson Correlation | -.051 | -.196 | 1 | .441" | .483" |
| | Slg. (2-talled) | .617 | .050 | . | .000 | .000 |
| | N | 100 | 100 | 100 | 100 | 100 |
| осведомленность | Pearson Correlation | -.134 | -.153 | .441** | 1 | .476" |
| | Slg. (2-talled) | .185 | .127 | ,000 | . | ,000 |
| | N | 100 | 100 | 100 | 100 | 100 |
| кратковременная память | Pearson Correlation | .017 | .016 | .483** | .476" | 1 |
| | Slg. (2-talled) | .867 | .886 | .000 | .000 | . |
| | N | 100 | 100 | 100 | 100 | 100 |

**. Correlation Is significant at the 0.01 level (2-talled).

**Рис. 9.4.** Фрагмент окна вывода после выполнения шага 5

После выполнения шага 5а с использованием редактора синтаксиса будет получена таблица результатов, представленная на рис. 9.5.

**Correlations**

| | | ОТМЕТКА2 |
|---|---|---|
| счет в уме | Pearson Correlation | .080 |
| | Sig. (2-talled) | .430 |
| | N | 100 |
| числовые ряды | Pearson Correlation | .114 |
| | Sig. (2-tailed) | .258 |
| | N | 100 |
| словарь | Pearson Correlation | .040 |
| | Slg. (2-talled) | .692 |
| | N | 100 |
| осведомленность | Pearson Correlation | .257 |
| | Sig. (2-tailed) | .010 |
| | N | 100 |
| кратковременная память | Pearson Correlation | .294 |
| | Slg. (2-talled) | .003 |
| | N | 100 |

**Рис.** 9.5. Фрагмент окна вывода после выполнения шага 5а

Значимая положительная корреляция в этой таблице наблюдается, в частности, между переменными тест5 и отметка2 ($r$ - 0,294, $p$ - 0,003). Это означает, что чем лучше кратковременная память (тест 5), тем выше средняя отметка за выпускной класс.

# 10 Средние значения



С помощью команды Crosstabs (Таблицы сопряженности), описанной в главе 8, вычисляются частоты по градациям неколичественных (номинативных) переменных. Таблицы сопряженности позволяют сравнивать частоты для разных подгрупп, которые соответствуют градациям номинативной переменной. Например, составив таблицу сопряженности полххобби, вы могли видеть, что среди девушек 15 увлекаются спортом, 27 — искусством и т. п. Команда Means (Средние) предназначена для сравнения подгрупп объектов по средним значениям количественных признаков. При этом предполагается, что в данных имеются не только количественные переменные, для которых вычисляются средние, но и номинативные переменные, разделяющие объекты на подгруппы. Команда Means (Средние) вполне применима и к данным файла ex01.sav, который мы рассматриваем в качестве примера. Так, при помощи этой команды можно сравнить средние значения успеваемости (отметка1, отметка2) юношей и девушек (пол), учащихся разных классов (класс) и т. д. Результаты вычислений *представляются* в виде таблиц, похожих на таблицы сопряженности при использовании команды Crosstabs (Таблицы сопряженности). Отличие заключается в том, что для каждой подгруппы вычисляется не трлько частота, но и среднее значение.

Команда Means (Средние) является одной из самых простых в SPSS. Для выбранных подгрупп она подсчитывает средние значения, стандартные отклонения и частоты. Кроме того, с помощью кнопки Options (Параметры) можно задать вывод результатов однофакторного дисперсионного анализа. В этой главе дисперсионный анализ упоминается без детального описания, поскольку этому вопросу целиком посвящены главы 13-16.

## Пошаговый алгоритм вычислений

В пошаговых процедурах используется файл ex01.sav с входящими в него переменными класс, пол и отметка2. Сначала выполняются три подготовительных шага.

**Шаг 1** Создайте новый файл данных или подготовьте существующий. Все необходимые действия описаны в главе 3.

**Шаг 2** Запустите программу SPSS при помощи значка на рабочем столе или команды Пуск ▸ Программы ▸ SPSS for Windows ▸ SPSS 11.5 for Windows (Start ▸ Programs ▸ SPSS for Windows ▸ SPSS 11.5 for Windows) главного меню Windows. В открывшемся после запуска программы диалоговом окне SPSS for Windows щелкните на кнопке Cancel (Отмена).

После выполнения этого шага на экране появится окно редактора данных SPSS.

**Шаг 3** Откройте файл данных, с которым вы намерены работать (в нашем случае это файл ex01.sav). Если он расположен в текущей папке, то выполните следующие действия:

1. Выберите в меню File (Файл) команду Open ▸ Data (Открытие ▸ Данные) или щелкните на кнопке Open File (Открытие файла) панели инструментов.
2. В открывшемся диалоговом окне дважды щелкните на имени ex01.sav или введите его с клавиатуры и щелкните на кнопке OK.

Независимо от того, открыта программа SPSS только что или какие-то процедуры уже выполнялись, в верхней части главного окна должна присутствовать строка меню (она показана ниже). Пока строка меню присутствует на экране, доступны все команды анализа данных. При этом окно с данными видеть не обязательно.

File Edit View Data Transform Analyze Graphs Utilities Window Help

При работе со сводными таблицами или при редактировании диаграмм некоторые пункты строки меню могут исчезать или меняться. Чтобы вернуться к главному окну и стандартной строке меню, щелкните на кнопке свертывания или восстановления текущего окна.

После завершения шага 3 на экране должно присутствовать окно редактора данных со строкой меню.

**Шаг 4** В меню Analyze (Анализ) выберите команду Compare Means > Means (Сравнение средних ▸ Средние). На экране появится диалоговое окно Means (Средние), показанное на рис. 10.1.

В диалоговом окне Means (Средние) вам необходимо задать переменные, которые будут участвовать в процедуре. Список Dependent List (Зависимые переменные) в верхней части окна предназначен для количественных переменных, характеристики которых будут вычисляться. Например, в качестве зависимых переменных могут выступать переменные отметка1, отметка2, тест1 и т. д. Список Dependent List (Зависимые переменные) может содержать несколько переменных, при этом для каждой из переменных можно задавать собственные наборы вычисляемых характеристик.

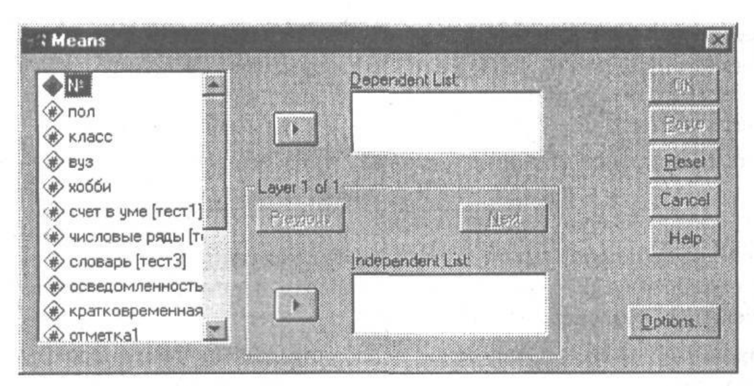



**Рис.** 10.1. Диалоговое окно Means

Список Independent List (Независимые переменные) служит для задания неколичественных (номинативных) переменных, градации которых определяют сравниваемые подгруппы объектов (пол, класс, вуз и т. п.). На основе этих переменных строится таблица сопряженности. Если включить в верхний список только переменную отметка2, а в нижний список — только переменную пол, то в результате выполнения команды Means (Средние) мы увидим средние значения отметок для юношей и девушек. По умолчанию для обеих величин также будут вычислены стандартные отклонения.

Как правило, исследователи не ограничиваются единственной независимой переменной. Типичным является построение таблицы сопряженности с участием двух переменных. Например, таблица пол×класс позволяет получить средние значения для юношей и девушек каждого из трех классов. Чтобы задать несколько переменных в списке Independent List (Независимые переменные), используйте кнопки Previous (Предыдущая) и Next (Следующая) справа и слева от метки Layer 1 of 1 (Слой 1 из 1). Как вы увидите в приведенных далее пошаговых процедурах, все действия в окне Means (Средние) просты и понятны.

**Шаг 5** После выполнения шага 4 должно быть открыто диалоговое окно Means (Средние), показанное на рис. 10.1. Мы вычислим средние значения отметок для учащихся каждого из трех классов.

1. Щелкните сначала на переменной отметка2, чтобы выделить ее, а затем - на верхней кнопке со стрелкой, чтобы переместить переменную в список Dependent List (Зависимые переменные).
2. Щелкните сначала на переменной класс, чтобы выделить ее, а затем — на нижней кнопке со стрелкой, чтобы переместить переменную в список Independent List (Независимые переменные).
3. Щелкните на кнопке OK, чтобы открыть окно вывода.

Можно включить в процедуру дополнительную независимую переменную, например пол, и тем самым построить таблицу сопряженности. Это делается следующим образом.

**Шаг 5а** После выполнения шага 4 должно быть открыто диалоговое окно Means (Средние), показанное на рис. 10.1. Если вы уже успели поработать с этим окном, очистите его щелчком на кнопке Reset (Сброс) и выполните следующие действия:

1. Щелкните сначала на переменной отметка2, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в список Dependent List (Зависимые переменные).
2. Щелкните сначала на переменной класс, чтобы выделить ее, а затем — на нижней кнопке со стрелкой, чтобы переместить переменную в список Independent List (Независимые переменные).
3. Щелчком на кнопке Next (Следующая) освободите список Independent List (Независимые переменные) — метка Layer 1 of 1 (Слой 1 из 1) изменится на Layer 2 of 2 (Слой 2 из 2). После этого щелкните сначала на переменной пол, чтобы выделить ее, а затем — на нижней кнопке со стрелкой, чтобы переместить переменную в список Independent List (Независимые переменные).
4. Щелкните на кнопке OK, чтобы открыть окно вывода результатов.

Количество зависимых переменных может быть любым. Программа просто строит в таблице столько столбцов, сколько переменных указано в списке Dependent List (Зависимые переменные).

В следующем примере мы вычислим средние значения и стандартные отклонения для переменных отметка1 и отметка2 с использованием таблицы сопряженности класс×пол.

**Шаг 5б** После выполнения шага 4 должно быть открыто диалоговое окно Means (Средние), показанное на рис. 10.1. Если вы уже успели поработать с этим окном, очистите его щелчком на кнопке Reset (Сброс) и выполните следующие действия:

1. Щелкните сначала на переменной отметка1, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в список Dependent List (Зависимые переменные).
2. Повторите предыдущее действие для переменной отметка2.
3. Щелкните сначала на переменной класс, чтобы выделить ее, а затем — на нижней кнопке со стрелкой, чтобы переместить переменную в список Independent List (Независимые переменные).
4. Щелчком на кнопке Next (Следующая) освободите список Independent List (Независимые переменные) — метка Layer 1 of 1 (Слой 1 из 1) изменится на Layer 2 of 2 (Слой 2 из 2). После этого щелкните сначала на переменной пол, чтобы выделить ее, а затем — на нижней кнопке со стрелкой, чтобы переместить переменную в список Independent List (Независимые переменные).
5. Щелкните на кнопке OK, чтобы открыть окно вывода результатов.

Для того чтобы выполнить однофакторный дисперсионный анализ или включить в ячейки дополнительную информацию, можно использовать кнопку Options

(Параметры) в диалоговом окне Means (Средние). После щелчка на этой кнопке откроется диалоговое окно Means: Options (Средние: Параметры), представленное на рис. 10.2.

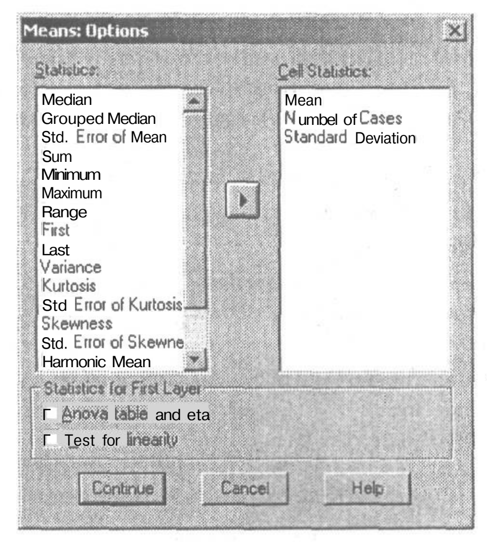



**Рис. 10.2.** Диалоговое окно Means: Options

С помощью диалогового окна Means: Options (Средние: Параметры) можно задать дополнительные параметры вывода для команды Means (Средние). Например, помимо величин, вычисляемых по умолчанию (среднего значения, стандартного отклонения и числа объектов), можно указать любую совокупность показателей, перечисленных в списке Statistics (Статистики). Для этого следует выделить нужный пункт списка, а затем щелчком па кнопке со стрелкой добавить его в список Cell Statistics (Статистики ячеек).

Как уже упоминалось, команда Means (Средние) позволяет выполнять однофакторный дисперсионный анализ. Для этого в группе Statistics for the First Layer (Статистики для первого слоя) нужно установить флажок ANOVA table and eta (Таблица ANOVA и коэффициент Эта). В процессе группировки зависимой переменной отметка2 по градациям независимой переменной класс программа путем однофакторного дисперсионного анализа сравнит три средних значения для градаций переменной класс.

В следующем примере мы вычислим средние значения для каждой из шести подгрупп, образующихся при сопряжении переменных класс и пол. Кроме того, мы проведем дисперсионный анализ с зависимой переменной отметка2 и независимой переменной класс.

После выполнения шага 5, 5а, 5б или 5в программа автоматически активизирует окно вывода результатов. Для просмотра результатов вы при необходимости

можете воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

**Шаг 5в** После выполнения шага 4 должно быть открыто диалоговое окно Means (Средние), показанное на рис. 10.1. Если вы уже успели поработать с этим окном, очистите его щелчком на кнопке Reset (Сброс) и выполните следующие действия:

1. Щелкните сначала на переменной отметка2, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в список Dependent List (Зависимые переменные).
2. Щелкните сначала на переменной класс, чтобы выделить ее, а затем — на нижней кнопке со стрелкой, чтобы переместить переменную в список Independent List (Независимые переменные).
3. Щелчком на кнопке Next (Следующая) освободите список Independent List (Независимые переменные) — метка Layer 1 of 1 (Слой 1 из 1) изменится на Layer 2 of 2 (Слой 2 из 2). После этого щелкните сначала па переменной пол, чтобы выделить ее, а затем — на нижней кнопке со стрелкой, чтобы переместить переменную в список Independent List (Независимые переменные).
4. Щелкните на кнопке Options (Параметры), чтобы открыть диалоговое окно Means: Options (Средние: Параметры), и установите флажок ANOVA table and Eta (Таблица ANOVA и коэффициент Эта).
5. Щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Means (Средние).
6. Щелкните на кнопке OK, чтобы открыть окно вывода.

# Печать результатов и выход из программы

Ниже описана типичная процедура печати результатов статистического анализа (или нескольких анализов). После выполнения шага 5 должно быть открыто окно вывода.

**Шаг 6** В окне вывода укажите фрагменты, выводимые на печать (см. раздел «Окно вывода» в главе 2), в меню File (Файл) выберите команду Print (Печать), при необходимости задайте параметры печати и щелкните на кнопке OK.

Последнее, что необходимо сделать после завершения исследования и печати результатов, — это выйти из программы SPSS.

**Шаг 7** Для выхода из программы в меню File (Файл) выберите команду Exit (Выход).

Иногда после выполнения команды Exit (Выход) на экране могут появляться небольшие диалоговые окна с вопросом о необходимости сохранения сделанных в файлах изменений и кнопками, описывающими возможные варианты ответа. Для завершения работы просто щелкайте на соответствующих кнопках.

# Представление результатов

На рис. 10.3 приведены фрагменты выводимых данных, сгенерированных программой после выполнения шага 5в.

**Report**

ОТМЕТКА2

| КЛАСС | ПОЛ | Mean | N | Std. Deviation |
|---|---|---|---|---|
| 1 | ЖЕН | 4.1286 | 14 | .21901 |
| | МУЖ | 4.0711 | 19 | .26579 |
| | Total | 4.0955 | 33 | .24506 |
| 2 | ЖЕН | 4.2048 | 21 | .20670 |
| | МУЖ | 4.1107 | 14 | .22717 |
| | Total | 4.1S71 | 35 | .21691 |
| 3 | ЖЕН | 4.4173 | 26 | .27711 |
| | МУЖ | 4.3667 | 6 | .26204 |
| | Total | 4.4078 | 32 | .27093 |
| Total | ЖЕН | 4.2779 | 61 | .26856 |
| | МУЖ | 4.1308 | 39 | .26622 |
| | Total | 4.2205 | 100 | .27589 |

**ANOVA Table**

| | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| ОТМЕТКА2 * КЛАСС | Between Groups | 1.738 | 2 | .869 | 14.54 | .000 |
| | Within Groups | 5.797 | 97 | .060 | | |
| | Total | 7.535 | 99 | | | |

**Measures of Association**

| | Eta | Eta Squared |
|---|---|---|
| ОТМЕТКА2 * КЛАСС | .480 | .231 |

**Рис. 10.3.** Фрагменты окна вывода после выполнения шага 5в

В первой таблице приводятся средние значения (Means), частоты (N) и стандартные отклонения (Std. Deviation) зависимой переменной отметка2 для каждой из подгрупп. Во второй и третьей таблицах содержатся результаты дисперсионного анализа. Обратите внимание, что дисперсионный анализ проведен только для средних значений первого слоя, который определяется градациями переменной класс. Для сравнения средних, соответствующих градациям переменной пол, следовало бы ее указать первой в списке Independent List (Независимые переменные). Более подробное описание однофакторного дисперсионного анализа приводится в главе 12, а трактовка остальных терминов окна вывода дана далее. Вычисленные средние значения (4,096, 4,167 и 4,408) различаются на уровне значимости $p = 0{,}000$. Это свидетельствует о статистически достоверной зависимости

успеваемости учащихся (отметка2) от класса, в котором они учатся (класс). Отметим, что команда Means (Средние) не позволяет определить статистическую значимость при попарном соотнесении средних значений. Так, в данном случае по результатам статистической проверки можно сделать вывод только о том, что классы различаются по успеваемости, без конкретизации, как эти классы отличаются друг от друга. Попарно сравнить средние значения можно при помощи $t$-критерия Стьюдента, применение которого рассматривается в следующей главе.

- Within Croups Sum of Squares (Внутригрупповая сумма квадратов) — сумма квадратов разностей между средним значением группы и каждым наблюдаемым значением этой группы.
- Between Croups Sum of Squares (Межгрупповая сумма квадратов) — сумма квадратов разностей между общим средним значением и средним значением каждой группы, умноженных на весовые коэффициенты, равные числу объектов в группе.
- Between Croups df (Межгрупповое число степеней свободы) — число групп, уменьшенное па единицу (3 - 1 = 2).
- Within Groups df (Внутригрупповое число степеней свободы) — разность между числом объектов и числом групп (105 - 3 = 102).
- Mean square (Средний квадрат) — отношение суммы квадратов к числу степеней свободы.
- F (F-критерий) — отношение межгруппового среднего квадрата к внутригрупповому среднему квадрату.
- Sig. (Значимость) — вероятность того, что наблюдаемые различия случайны. Величина $p$ = 0,000 свидетельствует о высокой статистической значимости различий.
- Eta (Коэффициент Эта) — мера связи между двумя переменными: количественной и номинативной.
- Eta Squared (Коэффициент Эта в квадрате) — мера влияния независимой переменной на дисперсию зависимой переменной. Величина 0,231 свидетельствует о том, что 23,1 % дисперсии зависимой переменной объясняются влиянием независимой переменной.

# 11 Сравнение двух средних и t-критерий



Различные варианты обработки данных с применением $t$-критерия позволяют сделать вывод о различии двух средних значений. Например, в случае $t$-критерия для независимых выборок проверяется достоверность различия двух выборок по переменной, измеренной у представителей этих двух выборок. Для этих выборок вычисляются средние значения, затем по $t$-критерию определяется статистическая значимость их различия. Применение $t$-критерия, по-видимому, самый распространенный метод статистического вывода, так как позволяет ответить на простой вопрос о различии двух выборок по уровню выраженности измеренного признака. Основное требование к данным для применения этого критерия — представление переменных, по которым сравниваются выборки, в метрической шкале измерения. Если есть сомнения в соответствии этому требованию, то можно воспользоваться непараметрическими методами, изложенными в главе 12.

SPSS позволяет применять 3 варианта $t$-критерия: Independent-Samples T Test (t-критерий для независимых выборок), Paired-Samples T Test (t-критерий для зависимых выборок), One-Samples T Test (t-критерий для одной выборки).

- Первый из вариантов $t$-критерия, $t$-критерий для *независимых выборок* (independent samples), предназначен для сравнения средних значений двух выборок. Для сравниваемых выборок должны быть определены значения одной и той же переменной. С помощью $t$-критерия для независимых выборок можно сравнить успеваемость студентов и студенток, степень удовлетворенности жизнью холостяков и женатых, средний рост футболистов двух команд и пр. Обязательным условием для проведения этого $t$-критерия является независимость выборок.

- Второй из $t$-критериев, $t$-критерий для *зависимых выборок* (paired samples, или dependent samples), позволяет сравнить средние значения двух измерений одного признака для одной и той же выборки, например результаты

первого и последнего экзаменов группы студентов или значения показателя до и после воздействия на группу. Обязательным условием применения $t$-критерия для зависимых выборок является наличие повторного измерения для одной выборки.

- Последний из $t$-критериев, $t$-критерий для *одной выборки* (one sample), позволяет сравнить среднее значение этой выборки с некоторой эталонной величиной. Например, отличается ли среднее значение некоторого теста для данной выборки от нормативной величины, отличается ли время, показанное бегунами во время соревнования, от 17 минут, и т. д.

# Уровень значимости

Результат сравнения средних значений с применением $t$-критерия оценивается по уровню значимости. Напомним, что уровень значимости является мерой статистической достоверности результата вычислений, в данном случае — различий средних, и служит основанием для интерпретации. Если исследование показало, что $p$-уровень значимости различий не превышает 0,05, это означает, что с вероятностью не более 5 % различия являются случайными. Обычно это является основанием для вывода о статистической достоверности различий. В противном случае ($p > 0{,}05$) различие признается статистически недостоверным и не подлежит содержательной интерпретации.

SPSS позволяет определять два уровня значимости: односторонний (one-tailed) и двусторонний (two-tailed). Обычно используется двусторонний уровень значимости. Но если вы заранее знаете направление различий и вас интересует только это направление, то можно использовать односторонний критерий значимости. Однако такая ситуация встречается редко, а если и встречается, то правомерность односторонней проверки с трудом поддается обоснованию.

Существенным различием между двумя типами уровней значимости является то, что двусторонний $p$-уровень оказывается вдвое больше, чем односторонний. По умолчанию SPSS вычисляет двусторонний уровень значимости, но вы легко можете получить односторонний уровень значимости, разделив вычисленное программой значение на 2.

# Пошаговые алгоритмы вычислений

Для применения $t$-критерия мы воспользуемся переменными из файла ex01.sav, но сначала необходимо выполнить три подготовительных шага.

Шаг 1 Создайте новый файл данных или подготовьте существующий. Все необходимые действия описаны в главе 3.

**Шаг 2** Запустите программу SPSS при помощи значка на рабочем столе или команды Пуск ▸ Программы ▸ SPSS for Windows ▸ SPSS 11.5 for Windows (Start ▸ Programs ▸ SPSS for Windows ▸ SPSS 11.5 for Windows) главного меню Windows. В открывшемся после запуска программы диалоговом окне SPSS for Windows щелкните на кнопке Cancel (Отмена).

После выполнения этого шага на экране появится окно редактора данных SPSS.

**Шаг 3** Откройте файл данных, с которым вы намерены работать (в нашем случае это файл ex01.sav). Если он расположен в текущей папке, то выполните следующие действия:

1. Выберите в меню File (Файл) команду Open ▸ Data (Открытие ▸ Данные) или щелкните на кнопке Open File (Открытие файла) панели инструментов.
2. В открывшемся диалоговом окне дважды щелкните па имени ex01.sav или введите его с клавиатуры и щелкните на кнопке OK.

Независимо от того, открыта программа SPSS только что или какие-то процедуры уже выполнялись, в верхней части главного окна должна присутствовать строка меню (она показана ниже). Пока строка меню присутствует на экране, доступны все команды анализа данных. При этом окно с данными видеть не обязательно.

File Edit View Data Transform Analyze Graphs Utilities Window Help

При работе с таблицами результатов или при редактировании диаграмм некоторые пункты строки меню могут исчезать или меняться. Чтобы вернуться к главному окну и стандартной строке меню, щелкните на кнопке свертывания или восстановления текущего окна.

## Применение t-критерия для независимых выборок

После завершения шага 3 на экране должно присутствовать окно редактора данных со строкой меню.

**Шаг 4** В меню Analyze (Анализ) выберите команду Compare Means ▸ Independent-Samples T Test (Сравнение средних ▸ t-критерий для независимых выборок). На экране появится диалоговое окно Independent-Samples T Test (t-критерий для независимых выборок), показанное на рис. 11.1.

В представленном диалоговом окне вы можете задать параметры для применения *t*-критерия. Левая часть окна содержит список всех доступных переменных файла, а правая часть — список Test Variable(s) (Тестируемые переменные), в который вам необходимо поместить имена переменных для применения *t*-критерия. Список может содержать любое количество переменных, а сами переменные должны быть метрического типа (это переменные отметка1, отметка2, тест1 и т. п.).

В поле Grouping Variable (Группирующая переменная) указывается имя переменной, градациям которой соответствуют две независимые выборки для $t$-критерия. Как правило, группирующая переменная дискретна и имеет 2 градации. В файле ex01.sav примером двухуровневой переменной является переменная пол. Группирующая переменная па самом деле может иметь более двух уровней и даже непрерывный тип, но в обоих случаях необходимо разбить множество значений переменной на две категории или задать две градации из существующих. Так, мы можем взять переменную класс в качестве группирующей переменной и указать, что в одну выборку следует включать все объекты, имеющие значение 1, а в другую выборку — объекты, имеющие значение 3. Это позволит сравнить средние значения для двух из трех классов.

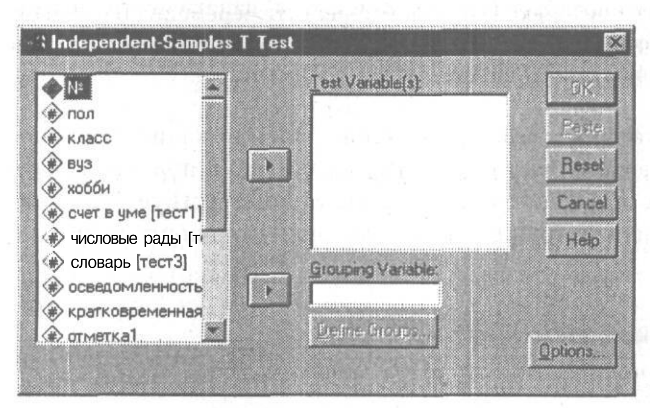



**Рис. 11.1.** Диалоговое окно Independent-Samples T Test

Итак, после того как имя группирующей переменной задано, необходимо задать ее категории. Это делается щелчком на кнопке Define Groups (Определение групп). Обратите внимание на то, что данный шаг необходим даже для двухуровневых переменных. На экране появится диалоговое окно Define Groups (Определение групп), показанное па рис. 11.2. Если установлен переключатель Use specified values (Использовать заданные значения), то в полях Group 1 (Группа 1) и Group 2 (Группа 2) нужно перечислить значения группирующей переменной, которые формируют первую и вторую выборки. Если установлен переключатель Cut Point (Точка деления), то в поле справа указывается значение, разбивающее весь диапазон переменной на два непересекающихся интервала.

В следующем примере в качестве группирующей мы используем переменную вуз. Она имеет 4 уровня, поэтому нам потребуется разбить эти уровни на 2 группы. Для этого мы установим в качестве точки раздела число 3. Это будет означать, что в первую выборку попадут учащиеся, выбирающие гуманитарные и экономические специальности, а во вторую — учащиеся, выбирающие технические и естественнонаучные специальности.

Рис. 11.2. Диалоговое окно Define Groups

**Шаг 5** На этом шаге мы применим $t$-критерий для сравнения юношей и девушек (градаций переменной пол) по переменной отметка2.

1. Щелкните сначала на переменной отметка2, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в список Test Variable(s) (Тестируемые переменные).
2. Щелкните сначала на переменной пол, чтобы выделить ее, а затем — на нижней кнопке со стрелкой, чтобы переместить переменную в поле Grouping Variable (Группирующая переменная).
3. Щелкните на кнопке Define Groups (Определение групп), чтобы открыть одноименное диалоговое окно, показанное на рис. 11.2.
4. Введите число 1 в поле Group 1 (Группа 1), нажмите клавишу Tab, введите число 2 в поле Group 2 (Группа 2) и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Independent-Samples T Test ($t$-критерий для независимых выборок).
5. Щелкните на кнопке OK, чтобы открыть окно вывода.

**Шаг 5а** После выполнения шага 4 должно быть открыто диалоговое окно Independent-Samples T Test (t-критерий для независимых выборок), показанное на рис. 11.1. Если вы уже успели поработать с этим окном, очистите его щелчком на кнопке Reset (Сброс) и выполните следующие действия:

1. Щелкните сначала на переменной отметка2, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в список Test Variable(s) (Тестируемые переменные).
2. Щелкните сначала на переменной вуз, чтобы выделить ее, а затем — на нижней кнопке со стрелкой, чтобы переместить переменную в поле Grouping Variable (Группирующая переменная).
3. Щелкните на кнопке Define Groups (Определение групп), чтобы открыть одноименное диалоговое окно, показанное на рис. 11.2.
4. Установите переключатель Cut point (Точка деления), введите в расположенное рядом поле значение 3 и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Independent-Samples T Test (t-критерий для независимых выборок).
5. Щелкните на кнопке OK, чтобы открыть окно вывода результатов.

## Применение t-критерия для зависимых выборок

Теперь мы рассмотрим два примера применения $t$-критерия для зависимых выборок. После завершения шага 3 на экране должно присутствовать окно редактора данных со строкой меню.

**Шаг 5б** В меню Analyze (Анализ) выберите команду Compare Means ▸ Paired-Samples T Test (Сравнение средних ▸ t-критерий для зависимых выборок). На экране появится диалоговое окно Paired-Samples T Test (t-критерий для зависимых выборок), показанное на рис. 11.3.

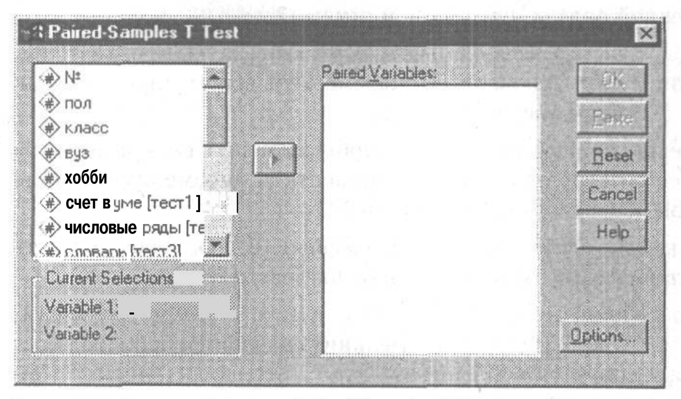



Рис. 11.3: Диалоговое окно Paired-Samples T Test

Применение $t$-критерия для зависимых выборок несколько проще, чем для независимых. Процедура управляется единственным диалоговым окном и не требует указания группирующей переменной. В левой части окна находится уже знакомый нам список доступных переменных файла, в котором вы можете выбрать переменные для применения $t$-критерия. Необычным здесь является то, что нужно указывать пары переменных — отсюда название целевого списка: Paired Variables (Пары переменных).

Сначала нужно щелкнуть на имени первой переменной, затем — на имени второй переменной, и после этого при помощи кнопки со стрелкой пара заносится в список. Количество выбираемых пар может быть любым.



**Шаг 5в** Чтобы сравнить отметки учащихся в 10 и 11 классах (отметка1 и отметка2), выполните следующие действия:

1. Щелкните на переменной отметка1. Ее имя появится в области Current Selections (Текущее выделение) рядом с меткой Variable 1 (Переменная 1).
2. Щелкните на переменной отметка2. Ее имя появится в области Current Selections (Текущее выделение) рядом с меткой Variable 2 (Переменная 2).
3. Щелкните на кнопке со стрелкой, чтобы переместить выбранную пару переменных в список Paired Variables (Пары переменных).
4. Щелкните на кнопке OK, чтобы открыть окно вывода.

Чтобы применить t-критерий одновременно для нескольких пар переменных, их все нужно переместить в список Paired Variables (Пары переменных). В следующем примере производится сравнение результатов первого теста (тест1) с результатами остальных четырех тестов.

**Шаг 5г** После выполнения шага 4а должно быть открыто диалоговое окно Paired-Samples T Test (t-критерий для зависимых выборок), показанное на рис. 11.3. Если вы уже успели поработать с этим окном, очистите его щелчком на кнопке Reset (Сброс) и выполните следующие действия:

1. Щелкните па переменной тест1. Ее имя появится в области Current Selections (Текущее выделение) рядом с меткой Variable 1 (Переменная 1).
2. Щелкните на переменной тест2. Ее имя появится в области Current Selections (Текущее выделение) рядом с меткой Variable 2 (Переменная 2).
3. Щелкните на кнопке со стрелкой, чтобы переместить выбранную пару переменных в список Paired Variables (Пары переменных).
4. Повторите три предыдущих действия, только вместо переменной тест2 во втором действии последовательно используйте переменные тест3, тест4 и тест5.
5. Щелкните на кнопке OK, чтобы открыть окно вывода.

## Применение t-критерия для одной выборки

Иногда бывает необходимо сравнить среднее значение распределения с какой-либо фиксированной величиной. Представим себе следующую ситуацию. Исследователь решил проверить, отличаются ли результаты тестирования его выборки от нормативных показателей. Предположим, нормативный показатель по тесту равен 10. Для того чтобы проверить результат выборки па соответствие норме, нужно вычислить среднее значение для выборки и сравнить его с числом 10. Подобные сравнения в программе SPSS реализуются при помощи *t*-критерия для одной выборки.

**Шаг 5д** В меню Analyze (Анализ) выберите команду Compare Means ▸ One-Sample T Test (Сравнение средних ▸ t-критерий для одной выборки). На экране появится диалоговое окно One-Sample T Test (t-критерий для одной выборки), показанное на рис. 11.4.

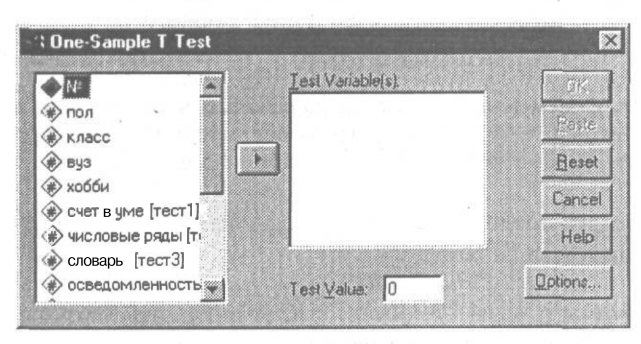



Рис. 11.4. Диалоговое окно One-Sample T Test

Применение $t$-критерия для одной выборки является самым простым с точки зрения как вычислений, так и управления. Все, что вам нужно сделать, — это выбрать одну или несколько переменных для анализа и, поместив ее (их) в список Test Variable(s) (Тестируемые переменные), задать константу для сравнения в поле Test Value (Тестовое значение), после чего щелкнуть на кнопке OK. Обратите внимание, что в случае задания в списке Test Variable(s) (Тестируемые переменные) нескольких переменных, их средние значения будут сравниваться с одной и той же величиной, указанной в поле Test Value (Тестовое значение). Если нужно провести сравнения с различными константами, необходимо применить $t$-критерий для каждой из констант.

**Шаг 5е** На этом шаге сравниваются средние значения переменных тест2 и тест3 с числом 10.

1. Щелкните сначала на переменной тест2, чтобы выделить ее, а затем — на кнопке со стрелкой, чтобы переместить переменную в список Test Variable(s) (Тестируемые переменные).
2. Повторите предыдущее действие для переменной тест3.
3. Нажмите клавишу Tab, чтобы перевести фокус ввода в поле Test Value (Тестовое значение), и введите число 10.
4. Щелкните на кнопке OK, чтобы открыть окно вывода.

После выполнения шага 5, 5а, 5б, 5в или 5г программа автоматически активизирует окно вывода. Для просмотра результатов вы при необходимости можете воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

# Печать результатов и выход из программы

Ниже описана типичная процедура печати результатов статистического анализа (или нескольких анализов). После выполнения шага 5 должно быть открыто окно вывода.

**Шаг 6** В окне вывода укажите фрагменты, выводимые на печать (см. раздел «Окно вывода» в главе 2), в меню File (Файл) выберите команду Print (Печать), при необходимости задайте параметры печати и щелкните на кнопке OK.

Последнее, что необходимо сделать после завершения исследования и печати результатов, — это выйти из программы SPSS.

**Шаг 7** Для выхода из программы в меню File (Файл) выберите команду Exit (Выход).

Иногда после выполнения команды Exit (Выход) на экране могут появляться небольшие диалоговые окна с вопросом о необходимости сохранения сделанных в файлах изменений и кнопками, описывающими возможные варианты ответа. Для завершения работы просто щелкайте на соответствующих кнопках.

# Представление результатов

В этом разделе мы приводим результаты применения всех трех видов *t*-критерия. После каждого из результатов следуют краткие пояснения, раскрывающие их смысл, а в конце раздела помещен список использовавшихся в окне вывода терминов. Визуальная форма вывода немного изменена, что никак не влияет на его содержимое.

## Результаты применения t-критерия для независимых выборок

На рис. 11.5 приведены фрагменты выводимых данных, сгенерированные программой после выполнения шага 5.

**Group Statistics**

| | ПОЛ | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| ОТМЕТКА2 | ЖЕН | 61 | 4.2779 | .26856 | .03439 |
| | МУЖ | 39 | 4.1308 | .26622 | .04263 |

**Independent Samples Test**

| | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | 95% Confidence Interval of the Difference | |
|---|---|---|---|---|---|---|---|---|---|
| | F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | Lower | Upper |
| Equal | .06 | .807 | 2.68 | 98 | .009 | .1471 | .05488 | .03820 | .25600 |
| Unequal | | | 2.69 | 81.6 | .009 | .1471 | .05477 | .03814 | .25606 |

Рис. 11.5. Фрагменты окна вывода после выполнения шага 5

Из результатов следует, что выборка из 39 юношей имеет средний балл 4,13, выборка из 61 девушки — средний балл 4,28. Различия статистически достоверны на высоком уровне значимости ($p$ - 0,009). Критерий равенства дисперсий Левина (Levene's Test) указывает на то, что дисперсии двух распределений статистически значимо не различаются ($p$ - 0,807). Этот результат позволяет применить несколько более мощный *t*-критерий для равных дисперсий (Equal). Если бы результаты критерия Левина были другими, то возникла бы необходимость в использовании *t*-критерия для неравных дисперсий (Unequal).

## Результаты применения t-критерия для зависимых выборок

На рис. 11.6 приведены фрагменты выводимых данных, сгенерированные программой после выполнения шага 5б.

**Paired Samples Statistics**

| | | Mean | N | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| Pair 1 | ОТМЕТКА1 | 3.9630 | 100 | .30597 | .03060 |
| | ОТМЕТКА2 | 4.2205 | 100 | .27589 | .02759 |

**Paired Samples Correlations**

| | | N | Correlation | Sig. |
|---|---|---|---|---|
| Pair 1 | ОТМЕТКА1 & ОТМЕТКА2 | 100 | .434 | .000 |

**Paired Samples Test**

| | Paired Differences | | | t | df | Sig. (2-tailed) |
|---|---|---|---|---|---|---|
| | | 95% Confidence Interval of the Difference | | | | |
| | Mean | Lower | Upper | | | |
| ОТМЕТКА1 - ОТМЕТКА2 | -.2575 | -.3191 | -.1959 | -8.29 | 99 | .000 |

**Рис. 11.6.** Фрагменты окна вывода после выполнения шага 5б

Как видно из результатов, для выборки объемом $N = 100$ среднее значение переменной отметка2 (4,22) оказалось статистически значимо выше среднего значения переменной отметка1 (3,96) с уровнем значимости $p < 0{,}001$. Кроме того, между переменными тест1 и тест2 существует значительная корреляция ($r = 0{,}434$, $p < 0{,}001$), свидетельствующая о том, что чем лучше результат обучения в классе 10, тем он лучше и в классе 11.

## Результаты применения t-критерия для одной выборки

На рис. 11.7 приведены фрагменты выводимых данных, сгенерированные программой после выполнения шага 5г.

Из таблиц видно, что среднее значение переменной тест2 (числовые ряды) составляет 10,35 и статистически достоверно не отличается от 10 ($p > 0{,}1$). Среднее значение переменной тест3 (словарь) равно 11,96 и статистически достоверно отличается от 10 ($p < 0{,}001$).

**One-Sample Statistics**

| | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|
| числовые ряды | 100 | 10.35 | 2.768 | .277 |
| словарь | 100 | 11.96 | 2.857 | .286 |

**One-Sample Test**

| | Test Value = 10 | | | | | |
|---|---|---|---|---|---|---|
| | t | df | Sig. (2-tailed) | Mean Difference | 95% Confidence Interval of the Difference | |
| | | | | | Lower | Upper |
| числовые ряды | 1.264 | 99 | .209 | .35 | -.20 | .90 |
| словарь | 6.861 | gg | .000 | 1.96 | 1.39 | 2.53 |

**Рис. 11.7.** Фрагменты окна вывода после выполнения шага 5г

# Терминология, используемая при выводе

Трактовка терминов, используемых программой в окне вывода результатов, дана далее.

- Std. Error (Стандартная ошибка) — отношение стандартного отклонения к квадратному корню из размера выборки *N*. Является мерой стабильности среднего значения.
- F (F-критерий) — величина, характеризующая соотношение дисперсий двух распределений.
- Sig. (Значимость) — в зависимости от того, равны или не равны дисперсии двух распределений, применяются различные виды статистических приближений. Величина $p > 0{,}05$ указывает на то, что дисперсии можно считать не различающимися.
- t ($t$-критерий) — $t$-критерий определяется как отношение разности средних значений к стандартному отклонению.
- df (Число степеней свободы) — для $t$-критерия с независимыми выборками при равенстве дисперсий число степеней свободы равно разности числа объектов и числа групп (100 - 2 = 98), а при различии дисперсий применяется более сложная формула, приводящая к дробному значению, равному 81,65. Для зависимых выборок и для одной выборки число степеней свободы для $t$-критерия определяется как 105 - 1 = 104.
- Sig. (2-Tailed) (Двусторонняя значимость) — по отношению к $t$-критерию двусторонняя значимость означает вероятность того, что разность между средними значениями является случайной, а по отношению к коэффициенту

корреляции — вероятность того, что связь между двумя переменными является случайной.

- Mean Difference (Разность средних) — разность между двумя средними значениями.
- Std. Deviation (Стандартное отклонение) — для $t$-критерия с зависимыми выборками это стандартное отклонение разности между значениями повторных измерений.
- Correlation (Корреляция) — мера связи двух переменных, а для зависимых выборок — мера связи повторных измерений. Численно определяется коэффициентом корреляции; в данном примере использовался коэффициент Пирсона. Более подробно корреляции описаны в главе 9.
- 95% Confidence Interval (Доверительный интервал в 95 %) — в случае $t$-критерия термин «доверительный интервал» относится к разности между средними значениями выборок.

# 12 Непараметрические критерии



Перед тем как ввести понятие непараметрического критерия, необходимо уточнить, что такое параметрический критерий. *Параметрический критерий* – это метод статистического вывода, который применяется в отношении *параметров* генеральной совокупности. Самым главным параметром для параметрических методов является нормальное распределение переменных и, как следствие, правомерность применения таких статистик, как среднее значение и стандартное отклонение. Несмотря на то что некоторые параметрические методы позволяют анализировать данные, распределенные по другим законам (например, биномиальному или Пуассона), непараметрические методы в этом смысле гораздо функциональнее, поскольку вообще не связывают анализ с каким-либо законом распределения.

Таким образом, непараметрические методы позволяют исследовать данные без каких-либо допущений о характере распределения переменных, в том числе при нарушении требования нормальности распределения. Так как эти методы предназначены для номинативных и ранговых переменных, в отношении которых недопустимо применение арифметических операций, они основаны на различных дополнительных вычислениях, среди которых можно отметить:

- ранжирование переменных;
- подсчет числа значений одного распределения, которые превышают значения другого распределения;
- применение весовых сравнений;
- определение степени отклонения распределения от случайного или биномиального распределения;
- проверка нормальности выборочного распределения;
- сравнения частот;

- сравнение групп путем вычисления частот значений, лежащих выше или ниже главной медианы.

Помимо всего прочего непараметрические критерии позволяют вычислять статистические показатели для одной выборки и сравнивать две выборки между собой. Несмотря на кажущуюся сложность непараметрические методы в большинстве своем очень просты для понимания и применения.

Для использования непараметрических методов мы задействуем уже знакомый нам файл ex01.sav и некоторые специально подготовленные примеры. Описание переменных файла ex01.sav можно найти в главе 3 (число объектов $N = 100$). Ссылки на имена файлов с дополнительными примерами будут даваться по мере изложения материала.

Структура этой главы несколько отличается от других. Из-за обилия методов пошаговые алгоритмы в этой главе объединены с описанием и интерпретацией результатов. После начальных трех шагов представлены два характерных для большинства непараметрических методов диалоговых окна. Далее даны описания восьми наиболее часто используемых непараметрических методов. Помимо общего описания каждого метода показаны пример его применения в виде шагов 4 и 5, 4а и 5а, 4б и 5б, и т. д., программный вывод результатов, его интерпретация и терминология. Указанные девять методов перечислены далее.

- Сравнение двух независимых выборок (*критерий Манна—Уитни*) позволяет установить различия между двумя независимыми выборками по уровню выраженности порядковой переменной.
- Сравнение двух зависимых выборок может проводиться по двум критериям. *Критерий знаков* основан на подсчете числа отрицательных и положительных разностей между повторными измерениями; *критерий Вилкоксона* в дополнение к знакам разностей учитывает их величину.
- *Критерий серий* определяет, является ли последовательность бинарных величин (событий) случайной или упорядоченной.
- *Биномиальный критерий* определяет, отличается ли распределение дихотомической величины от заданного соотношения.
- *Критерий Колмогорова—Смирнова для одной выборки* определяет отличие распределения переменной от нормального (равномерного, Пуассона и т. д.)
- *Критерий хи-квадрат для одной выборки* определяет степень отличия наблюдаемого распределения частот по градациям переменной от ожидаемого распределения.

- Сравнение *K* независимых выборок *(критерий Краскала—Уоллеса)* позволяет установить степень различия между тремя и более независимыми выборками по уровню выраженности порядковой переменной.
- Сравнение *K* зависимых выборок *(критерий Фридмана)* позволяет установить степень различия между тремя и более зависимыми выборками по уровню выраженности порядковой переменной.

# Пошаговые алгоритмы и результаты вычислений

При проведении непараметрических тестов сначала необходимо выполнить три подготовительных шага.

**Шаг 1** Создайте новый файл данных или подготовьте существующий. Все необходимые действия описаны в главе 3.

**Шаг 2** Запустите программу SPSS при помощи значка на рабочем столе или команды Пуск ▸ Программы ▸ SPSS for Windows ▸ SPSS 11.5 for Windows (Start ▸ Programs ▸ SPSS for Windows ▸ SPSS 11.5 for Windows) главного меню Windows. В открывшемся после запуска программы диалоговом окне SPSS for Windows щелкните на кнопке Cancel (Отмена).

После выполнения этого шага на экране появится окно редактора данных SPSS.

**Шаг 3** Откройте файл данных, с которым вы намерены работать (в нашем случае — это файл ex01.sav). Если он расположен в текущей папке, то выполните следующие действия:

1. Выберите в меню File (Файл) команду Open ▸ Data (Открытие ▸ Данные) или щелкните на кнопке Open File (Открытие файла) панели инструментов.
2. В открывшемся диалоговом окне дважды щелкните на имени ex01.sav или введите его с клавиатуры и щелкните на кнопке OK.

Независимо от того, открыта программа SPSS только что или какие-то процедуры уже выполнялись, в верхней части главного окна должна присутствовать строка меню (она показана ниже). Пока строка меню присутствует на экране, доступны все команды анализа данных. При этом окно с данными видеть не обязательно.



При работе с таблицами результатов или при редактировании диаграмм некоторые пункты строки меню могут исчезать или меняться. Чтобы вернуться к главному окну и стандартной строке меню, щелкните на кнопке свертывания или восстановления текущего окна.

## Некоторые диалоговые окна непараметрических тестов

Применение непараметрических методов иногда требует определять сравниваемые группы по градациям группирующей переменной. Для этого в соответствующих диалоговых окнах предусмотрена кнопка Define Groups (Определение групп), при щелчке на которой открывается характерное диалоговое окно, представленное на рис. 12.1. Если группирующая переменная имеет уровни 1 и 2, то в поле Croup 1 (Группа 1) следует ввести значение 1, в поле Group 2 (Группа 2) — значение 2, а затем щелкнуть на кнопке Continue (Продолжить).

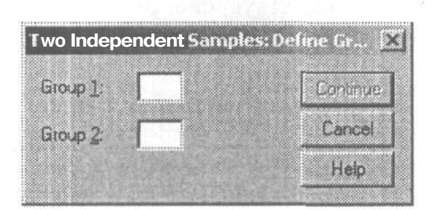



**Рис. 12.1.** Диалоговое окно определения групп

Еще одно новое для нас диалоговое окно появляется при щелчке па кнопке Options (Параметры) в основном диалоговом окне команды применения непараметрических методов (рис. 12.2). В группе Statistics (Статистика) имеются флажки Decriptives (Описательные статистики) и Quartiles (Квартили), позволяющие включить в выводимые данные соответственно описательные статистики (среднее значение, стандартное отклонение, минимум, максимум и объем выборки) и квартили. Группа переключателей Missing Values (Пропущенные значения) предназначена для выбора одного из двух режимов обработки отсутствующих значений (см. главу 4).

## Сравнение двух независимых выборок

*Критерий Манна—Уитни* (Mann—Whitney), или *U-критерий*, по назначению аналогичен $t$-критерию для независимых выборок. Разница заключается в том, что $t$-критерии ориентированы на нормальные и близкие к ним распределения, а критерий Манна—Уитни — на распределения, отличные от нормальных. В частном случае критерий Манна—Уитни можно применять и для нормально распределенных данных, однако он менее чувствителен к различиям (является менее мощным), чем $t$-критерий. Далее в примере мы попытаемся

выяснить, различаются ли юноши и девушки по успеваемости в выпускном классе. Напомним, что пол учащихся определяется значением неременной пол, а результат успеваемость — переменной отметка2. При реализации метода программа сначала ранжирует все объекты без учета принадлежности к сравниваемым группам, а затем вычисляет средние ранги для каждой из двух групп. Чем выше средний ранг группы, тем выше ее успеваемость. После определения средних рангов определяется $p$-уровень. Диалоговое окно $U$-критерия представлено на рис. 12.3.

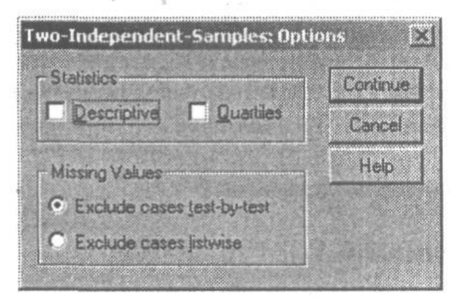



Рис. 12.2. Диалоговое окно настройки параметров непараметрических тестов

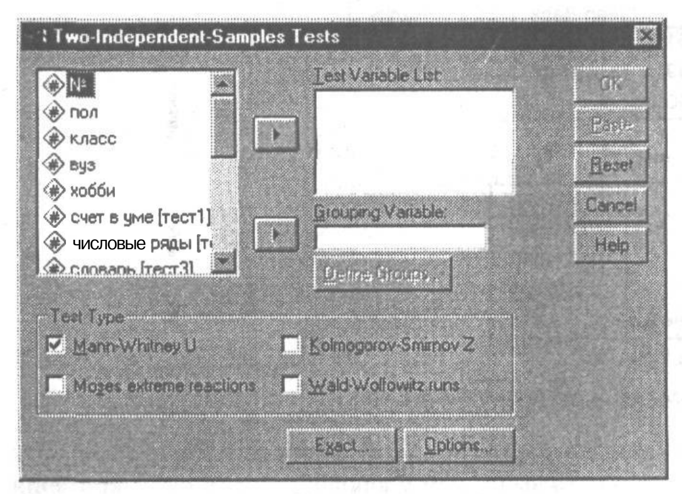



Рис. 12.3. Диалоговое окно U-критерия

После завершения шага 3 на экране должно присутствовать окно редактора данных. В этом примере мы сравним успеваемость юношей и девушек в 11 классе.

**Шаг 4** В меню Analyze (Анализ) выберите команду Nonparametric Tests ▸ 2 Independent Samples (Непараметрические методы ▸ Две независимые выборки), чтобы открыть диалоговое окно Two-Independent Samples Test (Критерий для двух независимых выборок), показанное на рис. 12.3.

**Шаг 5** Для применения метода выполните следующую последовательность действий:

1. Щелкните сначала на переменной отметка2, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в список Test Variable List (Список тестируемых переменных).
2. Щелкните сначала на переменной пол, чтобы выделить ее, а затем - на нижней кнопке со стрелкой, чтобы переместить переменную в поле Grouping Variable (Группирующая переменная).
3. Щелкните на кнопке Define Groups (Определение групп), чтобы открыть диалоговое окно, показанное на рис. 12.1.
4. В поле Group 1 (Группа 1) введите значение 1, нажмите клавишу Tab, чтобы переместить фокус ввода в поле Group 2 (Группа 2), введите значение 2 и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Two-Independent Samples Test (Критерий для двух независимых выборок).
5. Щелкните на кнопке OK, чтобы открыть окно вывода.

Результаты работы программы показаны на рис. 12.4.

**Ranks**

| | ПОЛ | N | Mean Rank | Sum of Ranks |
|---|---|---|---|---|
| ОТМЕТКА2 | ЖЕН | 61 | 56.21 | 3429.00 |
| | МУЖ | 39 | 41.56 | 1621.00 |
| | Total | 100 | | |

**Test Statistics[a]**

| | ОТМЕТКА2 |
|---|---|
| Mann-Whitney U | 841.000 |
| Wilcoxon W | 1621.000 |
| Z | -2.469 |
| Asymp. Sig. (2-tailed) | .014 |

a. Grouping Variable: ПОЛ

Рис. 12.4. Фрагменты окна вывода после выполнения шага 5

Средний ранг (Mean Rank) для девушек равен 56,21, а для мужчин — 41,56. Это значит, что у девушек успеваемость выше, чем у юношей. Величина *U*-критерия (Mann—Whitney U) равна 841. Значение Z является нормализованным, связанным с уровнем значимости $p = 0{,}014$. Поскольку величина уровня значимости (Asymp. Sig. (2-tailed)) меньше 0,05, мы можем быть уверены в достоверности вывода о том, что успеваемость девушек действительно выше успеваемости юношей.

## Сравнение двух зависимых выборок

Основные методы, которые используются для сравнения двух зависимых выборок, — это *критерий знаков* (sign test) и *критерий Вилкоксона* (Wilcoxon test).

## Критерий знаков

Критерий знаков позволяет сравнить два измерения переменной на одной выборке (например, «до» и «после») по уровню ее выраженности путем сопоставления количества положительных и отрицательных разностей (сдвигов) значений. Для того чтобы продемонстрировать применение критерия, сравним результаты учащихся по второму (тест2) и четвертому (тест4) тестам. Для каждого объекта сначала определяется знак разности значений. Так, для первого объекта значение тест2 равно 7, а значение тест4 — 10. Сравнение этих значений даст отрицательный знак (7 - 10 - 3 < 0). Для четвертого объекта значения переменных тест2 и тест4 составляют, соответственно, 9 и 6, и знак разности будет положительным (9 - 6 > 0). Подсчитывается число положительных, отрицательных и нулевых разностей, а затем вычисляется нормализованное *z*-значение и *p*-уровень значимости. Основное диалоговое окно критерия знаков представлено на рис. 12.5.

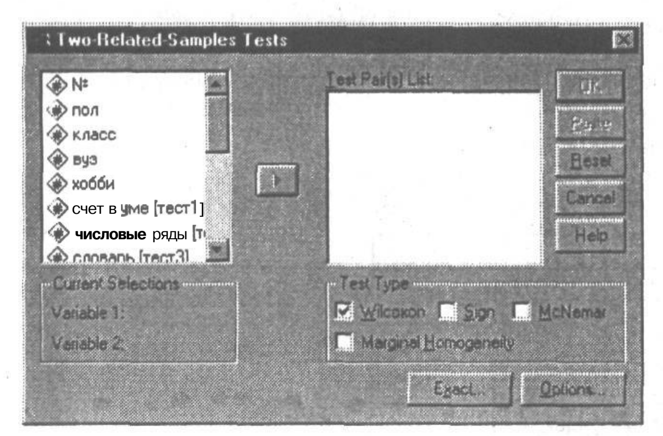



**Рис. 12.5.** Диалоговое окно для сравнения двух зависимых выборок

После завершения шага 3 на экране должно присутствовать окно редактора данных. В этом примере мы сравним результаты учащихся по второму (тест2) и четвертому (тест4) тестам.

**Шаг 4а** В меню Analyze (Анализ) выберите команду Nonparametric Tests ▸ 2 Related Samples (Непараметрические критерии ▸ Две зависимые выборки), чтобы открыть диалоговое окно Two-Related Samples Test (Критерий для двух зависимых выборок), показанное на рис. 12.5.

Результаты работы программы показаны на рис. 12.6.

Полученные результаты говорят о том, что в 39 случаях значения переменной тест2 оказались меньшими, чем значения переменной тест4, в 57 случаях значения переменной тест2 превысили значения переменной тест4 и 4 раза было установлено равенство значений обеих переменных. Стандартизованное значение составляет -1,735, а уровень значимости *p* - 0,083. Это означает, что различия между

результатами тестов тест1 и тест2 статистически недостоверны. Обратите внимание, поскольку переменные тест1 и тест2 являются метрическими, к ним предпочтительней применить $t$-критерий для зависимых выборок. Он показал бы, что средние значения тест1 и тест2 различаются с уровнем значимости $p = 0{,}01$. Таким образом, можно на практике убедиться в том, что статистические возможности $t$-критерия в отношении переменных значительно выше, чем возможности критерия знаков.

**Шаг 5а** Для применения критерия знаков выполните следующую последовательность действий:

1. В группе Test Type (Тип критерия) установите флажок Sign (Знаков) и сбросьте флажок Wilcoxon (Вилкоксона).
2. Щелкните на переменной тест2. Ее имя окажется возле метки Variable 1 (Переменная 1) в области Current Selections (Текущие выделения).
3. Нажмите клавишу Ctrl и щелкните на переменной тест4. Ее имя окажется возле метки Variable 2 (Переменная 2) в области Current Selections (Текущие выделения).
4. Щелкните на кнопке со стрелкой, чтобы переместить выбранную пару переменных в список Test Pair(s) List (Список тестируемых пар).
5. Щелкните на кнопке OK, чтобы открыть окно вывода.

**Frequencies**

| | | N |
|---|---|---|
| осведомленность - числовые ряды | Negative Differences[a] | 39 |
| | Positive Differences[b] | 57 |
| | Ties[c] | 4 |
| | Total | 100 |

a. осведомленность < числовые ряды

b. осведомленность > числовые ряды

c. осведомленность = числовые ряды

**Test Statistics[a]**

| | осведомленность - числовые ряды |
|---|---|
| Z | -1.735 |
| Asymp. Sig. (2-tailed) | .083 |

a. Sign Test

**Рис. 12.6.** Фрагменты окна вывода после выполнения шага 5а

## Критерий Вилкоксона

Недостатком критерия знаков является то, что он никак не учитывает величину разности двух значений. Так, для него нет разницы между результатами сравнения пар (10; 0) и (6; 5): в обоих случаях знак разности будет положительным. Однако очевидно, что абсолютное значение разности также характеризует соотно-

шение распределений. Для того чтобы учесть это, применяется критерий Вилкоксона. Этот критерий основан на подсчете абсолютных разностей между парами значений с последующим их ранжированием. Затем вычисляются средние значения рангов для положительных и отрицательных разностей (сдвигов). Уровень значимости подсчитывается на основе стандартизованного значения. Основное диалоговое окно критерия Вилкоксона то же, что и для критерия знаков (см. рис. 12.5).

После завершения шага 3 на экране должно присутствовать окно редактора данных. В этом примере мы, как и в предыдущем, сравним результаты учащихся по второму (тест2) и четвертому (тест4) тестам. Однако на этот раз вместо критерия знаков воспользуемся критерием Вилкоксона.

**Шаг 4б** В меню Analyze (Анализ) выберите команду Nonparametric Tests ▸ 2 Related Samples (Непараметрические критерии ▸ Две зависимые выборки), чтобы открыть диалоговое окно Two-Related Samples Test (Критерий для двух зависимых выборок), показанное на рис. 12.5. Если вы уже успели поработать с этим окном, щелкните на кнопке Reset (Сброс).

**Шаг 5б** Для применения критерия Вилкоксона выполните следующую последовательность действий:

1. Щелкните на переменной тест2. Ее имя окажется возле метки Variable 1 (Переменная 1) в области Current Selections (Текущие выделения).
2. Нажмите клавишу Ctrl и щелкните на переменной тест4. Ее имя окажется возле метки Variable 2 (Переменная 2) в области Current Selections (Текущие выделения).
3. Щелкните на кнопке со стрелкой, чтобы переместить выбранную пару переменных в список Test Pair(s) List (Список тестируемых пар).
4. Щелкните на кнопке OK, чтобы открыть окно вывода.

Обратите внимание, что мы не устанавливали флажок Wilcoxon (Вилкоксон), поскольку он установлен по умолчанию. Результаты работы программы показаны на рис. 12.7.

Результаты применения критерия Вилкоксона и критерия знаков очень похожи. Частота каждого из трех исходов $N$ осталась неизменной. Информация о каждом из исходов (кроме равенства) теперь включает также среднее и суммарное значения для соответствующих рангов. Визуальный анализ исходных данных говорит о том, что значения теста 4 (осведомленность) в целом несколько превышают значения теста 2 (числовые ряды). Это демонстрирует и величина $Z = -2{,}612$, которая значительно превосходит по модулю соответствующее значение, полученное ранее для критерия знаков. Уровень значимости $p = 0{,}013$; это значение говорит о статистической достоверности различий. Таким образом, мы убеждаемся в том, что критерий Вилкоксона является более чувствительным к различиям (более мощным), чем критерий знаков. Тем не менее он оказывается несколько хуже $t$-критерия, обеспечивающего уровень значимости 0,01, что подтверждает предпочтительность последнего для анализа метрических данных.

**Ranks**

| | | N | Mean Rank | Sum of Ranks |
|---|---|---|---|---|
| осведомленность - числовые ряды | Negative Ranks | 39[a] | 42.26 | 1648.00 |
| | Positive Ranks | 57[b] | 52.77 | 3008.00 |
| | Ties | 4[c] | | |
| | Total | 100 | | |

a. осведомленность < числовые ряды

b. осведомленность > числовые ряды

c. осведомленность = числовые ряды

**Test Statistics[b]**

| | осведомленность - числовые ряды |
|---|---|
| Z | -2.493[a] |
| Asymp. Sig. (2-tailed) | .013 |

a. Based on negative ranks.

b. Wilcoxon Signed Ranks Test

Рис. 12.7. Фрагменты окна вывода после выполнения шага 5б

## Критерий серий

Как следует из названия, *критерий серий* применяется для анализа последовательности объектов (явлений, событий), упорядоченных во времени или в порядке возрастания (убывания) значений измеренного признака. Кроме того, критерий требует представления последовательности в виде бинарной переменной, то есть как чередования событий 0 и 1. Математическая идея критерия основана на подсчете числа серий в упорядоченной последовательности событий двух типов, например, 0 и 1. *Серия* — это последовательность однотипных событий, непосредственно перед и после которой произошли события другого типа. Гипотеза о случайном распределении событий 1 среди событий 0 может быть отклонена, если количество серий либо слишком мало (однотипные события имеют тенденцию к группированию), либо слишком велико (события 0 и 1 имеют тенденцию к чередованию).

Основное диалоговое окно критерия серий представлено на рис. 12.8.

После завершения шага 3 на экране должно присутствовать окно редактора данных. Рассмотрим применение критерия серий на примере проверки гипотезы о неслучайном чередовании юношей и девушек (переменная пол) в списке испытуемых в файле ex01 .sav. Для задания точки раздела значений переменной на 2 категории предназначены флажок и поле Custom (Настройка). В данном случае в одну группу мы включим значения переменной пол, равные 1, а в другую группу — значения, равные 2.

**Шаг 4в** 5 В меню Analyze (Анализ) выберите команду Nonparametric Tests ▸ Runs (Непараметрические критерии ▸ Серии), чтобы открыть диалоговое окно Runs Test (Критерий серий), показанное на рис. 12.8:

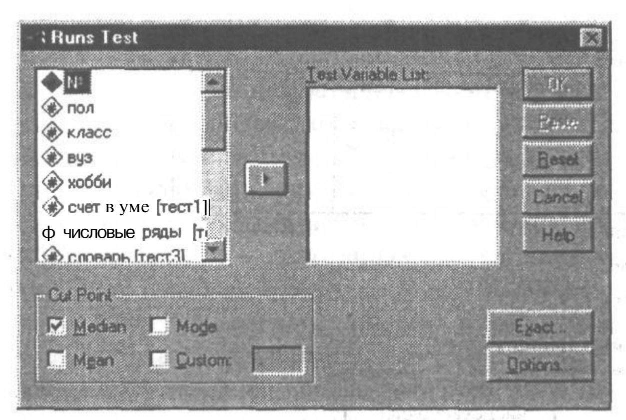



Рис. 12.8. Диалоговое окно критерия серий

**Шаг 5в** Для применения критерия выполните следующую последовательность действий:

1. Щелкните сначала на переменной пол, чтобы выделить ее, а затем — на кнопке со стрелкой, чтобы переместить переменную в поле Test Variable List (Список тестируемых переменных).
2. В группе Cut Point (Точка раздела) установите флажок Custom (Настройка), введите в расположенное рядом поле значение 2 и сбросьте флажок Median (Медиана).
3. Щелкните на кнопке OK, чтобы открыть окно вывода.

Результаты работы программы показаны на рис. 12.9.

**Runs Test**

| | Test Value[a] | Total Cases | Number of Runs | Z | Asymp. Sig. (2-tailed) |
|---|---|---|---|---|---|
| ПОЛ | 2 | 100 | 49 | .089 | .929 |

a. User-specified.

Рис. 12.9. Фрагмент окна вывода после выполнения шага 5в

Количество серий равно 49. В результаты включено значение точки деления, введенное в поле Custom (Настройка). Величина Z и соответствующая значимость зависят от числа серий. Число серий преобразуется к $z$-значению, для которого и определяется $p$-уровень. Большое значение $p$-уровня (0,929) свидетельствует о том, что чередование юношей и девушек в файле ex01.sav является случайным. Статистически значимый результат свидетельствовал бы о том, что чередование юношей и девушек в файле является неслучайным. Если при этом число серий было бы слишком велико, это свидетельствовало бы о том, что после юноши с высокой долей вероятности следует девушка (и наоборот). При малом значении числа серий можно было бы сделать вывод о том, что более вероятно группирование испытуемых в списке по половому признаку (после юноши чаще следует юноша, а после девушки — девушка).

## Биномиальный критерий

Назначение *биномиального критерия* — определение вероятности того, что наблюдаемое распределение не отличается от ожидаемого (заданного) биномиального распределения. Свойством биномиального распределения является заранее заданное соотношение вероятностей двух взаимоисключающих событий (обычно — равновероятное). Например, при многократном подбрасывании «правильной» монеты распределение вероятностей выпадения «орлов» и «решек» подчиняется биномиальному распределению. В качестве примера мы снова исследуем распределение юношей и девушек в файле ex01.sav. Зная заранее, что в файле собраны данные о 61 ученице и 41 ученике, мы сможем проверить, отличается ли статистически достоверно это распределение (наблюдаемое) от ожидаемого (теоретического) равновероятного соотношения. Основное диалоговое окно биномиального теста представлено на рис. 12.10.

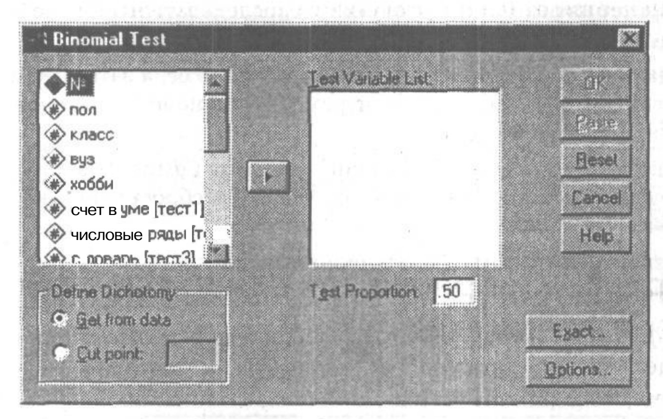



**Рис. 12.10.** Диалоговое окно биномиального критерия

После завершения шага 3 на экране должно присутствовать окно редактора данных. В этом примере мы попытаемся выяснить, является ли распределение мужчин и женщин в выборке биномиальным.

**Шаг 4г** В меню Analyze (Анализ) выберите команду Nonparametric Tests ▸ Binomial (Непараметрические критерии ▸ Биномиальный критерий), чтобы открыть диалоговое окно Binomial Test (Биномиальный критерий), показанное на рис. 12.6.

**Шаг 5г** Для применения биномиального критерия выполните следующую последовательность действий:

1. Щелкните сначала на переменной пол, чтобы выделить ее, а затем — на кнопке со стрелкой, чтобы переместить переменную в список Test Variable List (Список тестируемых переменных). Обратите внимание, что ожидаемое соотношение, заданное по умолчанию (Test Proportion) равно 0,5.
2. Щелкните на кнопке OK, чтобы открыть окно вывода.

Результаты работы программы показаны на рис. 12.11.

**Binomial Test**

| | | Category | N | Observed Prop. | Test Prop. | Asymp. Sig. (2-tailed) |
|---|---|---|---|---|---|---|
| ПОЛ | Group 1 | МУЖ | 39 | .39 | .50 | .035[a] |
| | Group 2 | ЖЕН | 61 | .61 | | |
| | Total | | 100 | 1.00 | | |

a. Based on Z Approximation.

**Рис. 12.11.** Фрагмент окна вывода после выполнения шага 5г

Ожидаемая пропорция для биномиального теста равна 0,5 для обеих групп. Наблюдаемая пропорция для каждой из групп определяется как отношение размера группы ($N$) к размеру выборки (100). Как можно видеть, наблюдаемые пропорции значительно отличаются от 0,5 и составляют 0,39 для мужчин и 0,61 для женщин. Уровень значимости, равный 0,035, свидетельствует о статистически достоверном отличии исследуемого распределения от биномиального (равновероятного).

При желании можно проверять отличие наблюдаемого распределения от любого другого биномиального распределения, задавая необходимые ожидаемые пропорции в поле Test Proportion (Ожидаемое соотношение).

## Критерий Колмогорова—Смирнова для одной выборки

*Критерий Колмогорова—Смирнова для одной выборки* позволяет определить, отличается ли заданное распределение от нормального (эксцесс и асимметрия распределения равны 0), равномерного (значения распределены с одинаковой плотностью, например, как у целых чисел от 1 до 1000), Пуассона (среднее значение и дисперсия равны $\lambda$; при больших значениях $\lambda$ распределение Пуассона приближается к нормальному) или экспоненциального. Суть метода заключается в сравнении эмпирического (наблюдаемого) распределения накопленных частот выборки с теоретическим (ожидаемым) распределением накопленных частот (нормальным, Пуассона и т. д.). В качестве примера мы исследуем распределение значений переменной отметка1 в файле ex01.sav на соответствие нормальному распределению. Основное диалоговое окно критерия Колмогорова—Смирнова для одной выборки представлено на рис. 12.12.

После завершения шага 3 на экране должно присутствовать окно редактора данных. В этом примере мы сравним распределение переменной отметка1 с нормальным распределением.

**Шаг 4д** В меню Analyze (Анализ) выберите команду Nonparametric Tests ▸ 1-Sample K-S (Непараметрические критерии ▸ Критерий K-C для одной выборки), чтобы открыть диалоговое окно One-Sample Kolmogorov—Smirnov Test (Критерий Колмогорова—Смирнова для одной выборки), показанное на рис. 12.12.

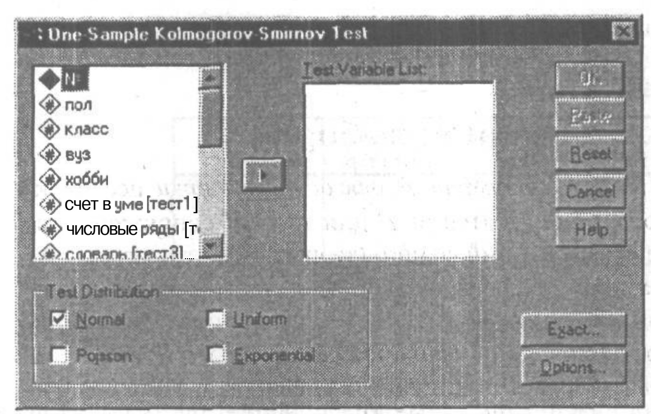



**Рис. 12.12.** Диалоговое окно критерия Колмогорова—Смирнова для одной выборки

**Шаг 5д** Для применения критерия выполните следующую последовательность действий:

1. Щелкните сначала на переменной отметка1, чтобы выделить ее, а затем — на кнопке со стрелкой, чтобы переместить переменную в список Test Variable List (Список тестируемых переменных).
2. Щелкните на кнопке OK, чтобы открыть окно вывода.

Результаты работы программы показаны на рис. 12.13.

**One-Sample Kolmogorov-Smirnov Test**

| | | ОТМЕТКА1 |
|---|---|---|
| N | | 100 |
| Normal Parameters [a,b] | Mean | 3.9630 |
| | Std. Deviation | .30597 |
| Most Extreme Differences | Absolute | .072 |
| | Positive | .072 |
| | Negative | -.051 |
| Kolmogorov-Smirnov Z | | .716 |
| Asymp. Sig. (2-tailed) | | .685 |

a. Test distribution is Normal.

b. Calculated from data.

**Рис. 12.13.** Фрагмент окна вывода после выполнения шага 5д

В первой строке таблицы приведен объем выборки (N); далее следуют ее среднее значение и стандартное отклонение. В строке Most Extreme Differences приведены наибольшие по абсолютному значению (Absolute), а также положительные (Positive) и отрицательные (Negative) отклонения исследуемого распределения от теоретического (в данном случае нормального). Строка Kolmogorov—Smirnov Z содержит $z$-значение, уровень значимости которого равен 0,685 (последняя

строка). Это означает, что распределение значений переменной отметка1 статистически не отличается от нормального.

## Критерий хи-квадрат для одной выборки

Рассматриваемый в данном разделе *критерий $\chi^2$ для одной выборки* несколько отличается от рассмотренного ранее критерия $x^2$ для таблиц сопряженности (см. главу 8). В данном случае в качестве ожидаемого (теоретического) распределения обычно выступает равномерное распределение объектов по градациям переменной, в отношении которой применяется критерий. Далее будет приведен пример применения критерия $\chi^2$ к переменной вуз из файла exQ1 .sav. Поскольку число объектов (*N*) в файле ex01.sav равно 100, а переменная вуз имеет 4 градации, ожидаемые частоты для каждой градации равны 100/4 = 25. Применение рассматриваемого критерия допускает задание не только равномерного ожидаемого распределения, но и любого другого. Например, можно проверить гипотезу о том, что соотношение учащихся, предпочитающих 4 категории специализаций, соотносятся как 20:20:30:30. Для этого в группе Expected Values (Ожидаемые значения) следует установить переключатель Values (Значения), а затем при помощи поля и кнопки Add (Добавление) последовательно ввести в список значения 20, 20, 30, 30. После этих действий ожидаемые частоты изменятся в соответствии с заданными пропорциями. Основное диалоговое окно критерия $\chi^2$ для одной выборки представлено на рис. 12.14.

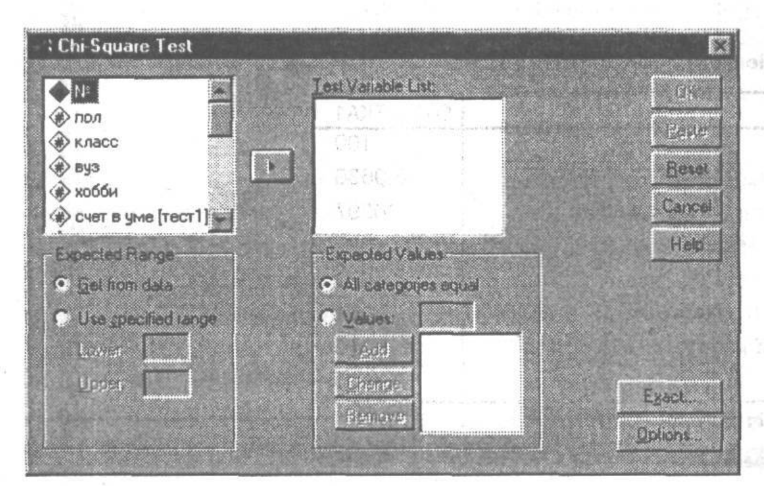



**Рис. 12.14.** Диалоговое окно критерия $\chi^2$ для одной выборки

После завершения шага 3 па экране должно присутствовать окно редактора данных. В этом примере мы применим критерий $\chi^2$ к переменной вуз.

**Шаг 4е** В меню Analyze (Анализ) выберите команду Nonparametric Tests ▸ Chi-Square (Непараметрические критерии ▸ Хи-квадрат), чтобы открыть диалоговое окно Chi-Square Test (Критерий хи-квадрат), показанное на рис. 12.14.

**Шаг 5е** Для применения критерия выполните следующую последовательность действий:

1. Щелкните сначала на переменной вуз, чтобы выделить ее, а затем — на кнопке со стрелкой, чтобы переместить переменную в список Test Variable List (Список тестируемых переменных).
2. Щелкните на кнопке (Ж, чтобы открыть окно вывода.

Результаты работы программы показаны па рис. 12.15.

**ВУЗ**

| | Observed N | Expected N | Residual |
|---|---|---|---|
| ГУМ | 22 | 25.0 | -3.0 |
| ЭКОН | 31 | 25.0 | 6.0 |
| ТЕХН | 11 | 25.0 | -14.0 |
| ЕСТ_Н | 36 | 25.0 | 11.0 |
| Total | 100 | | |

**Test Statistics**

| | ВУЗ |
|---|---|
| Chi-Square[a] | 14.480 |
| df | 3 |
| Asymp. Sig. | .002 |

[a] 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 25.0.

**Рис. 12.15.** Фрагменты окна вывода после выполнения шага 5е

Первая из таблиц демонстрирует заметные различия наблюдаемых и ожидаемых частот. Разность между наблюдаемыми и ожидаемыми частотами называется остатком (Residual). Число степеней свободы (df) определяется как число градаций переменной, уменьшенное на 1. Уровень значимости ($p$= 0,02) свидетельствует о статистически достоверном отличии наблюдаемого распределения предпочтений от равномерного распределения.

## Сравнение К независимых выборок и критерий Краскала—Уоллеса

Для сравнения более двух независимых выборок но уровню выраженности переменной применяется несколько критериев: *Н-критерий Краскала—Уоллеса* (Kruskal—Wallis Я), *критерий медианы* (median), *критерий Джонкира—Терпстра* (Jonckheere—Terpstra). Из них наибольшей чувствительностью к различиям обладает Я-критерий Краскала—Уоллеса. Этот критерий является непараметрическим аналогом однофакторного дисперсионного анализа, отличаясь от него в двух отношениях. Во-первых, критерий Краскала—Уоллеса основан не на сравнении

средних значений и дисперсий переменных, а на сравнении средних рангов. Во-вторых, вместо вычисления $F$-критерия на основе сравнения средних рангов с ожидаемыми значениями вычисляется критерий хи-квадрат. Для нормальных распределений однофакторный дисперсионный анализ обеспечивает более точные результаты, чем критерий Краскала—Уоллеса, однако применение последнего рекомендуется для распределений, отличающихся от нормального.

$H$-критерий Краскела—Уоллеса «по идее» сходен с $U$-критерием Манна—Уитни. Как и последний, он оценивает степень пересечения (совпадения) нескольких рядов значений измеренного признака. Чем меньше совпадений, тем больше различаются ряды, соответствующие сравниваемым выборкам. Основная идея $H$-критерия Краскела—Уоллеса основана па представлении всех значений сравниваемых выборок в виде одной общей последовательности упорядоченных (ранжированных) значений с последующим вычислением среднего ранга для каждой из выборок. Если выполняется статистическая гипотеза об отсутствии различий, то можно ожидать, что все средние ранги примерно равны и близки к общему среднему рангу.

Диалоговое окно непараметрических критериев для нескольких независимых выборок представлено на рис. 12.16.

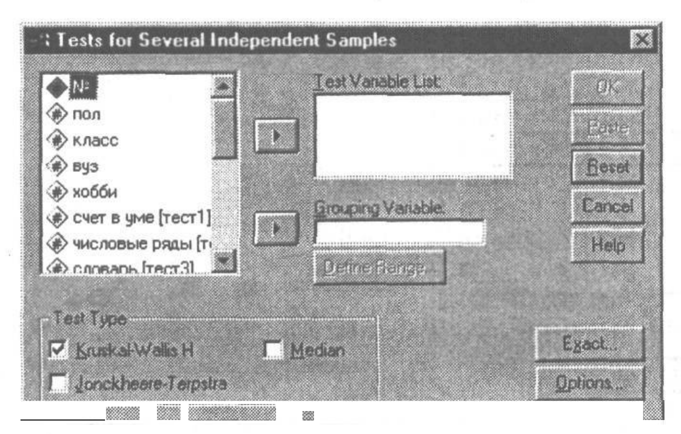



**Рис. 12.16.** Диалоговое окно критериев для нескольких независимых выборок

После завершения шага 3 на экране должно присутствовать окно редактора данных. В этом примере мы проведем сравнение трех групп учащихся, отличающихся внешкольными увлечениями (переменная хобби) и успеваемостью в 11 классе (переменная отметка2).

**Шаг 5ж** В меню Analyze (Анализ) выберите команду Nonparametric Tests ▸ K Independent Samples (Непараметрические критерии ▸ K независимых выборок), чтобы открыть диалоговое окно Test for Several Independent Samples (Критерии для нескольких независимых выборок), показанное на рис. 12.16.

**Шаг 5з** Для применения критерия выполните следующую последовательность действий:

1. Щелкните сначала на переменной отметка2, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в список Test Variable List (Список тестируемых переменных).
5 2. Щелкните сначала па переменной хобби, чтобы выделить ее, а затем — на нижней кнопке со стрелкой, чтобы переместить переменную в поле Grouping Variable (Группирующая переменная).
3. Щелкните на кнопке Define Range (Определение диапазона), чтобы открыть одноименное диалоговое окно.
4. В поле Minimum (Минимум) введите значение 1, нажмите клавишу Tab, чтобы переместить фокус ввода в поле Maximum (Максимум), введите значение 3 и щелкните на кнопке Continue (Продолжить), чтобы вернуться диалоговое окно Test for Several Independent Samples (Критерии для нескольких независимых выборок).
5. Щелкните на кнопке OK, чтобы открыть окно вывода.

Обратите внимание, что флажок Kruskal—Wallis H (Я-критерий Краскала—Уоллеса) установлен по умолчанию, поэтому мы ничего не меняли для выбора критерия.

Результаты работы программы показаны на рис. 12.17.

**Ranks**

| | ХОББИ | N | Mean Rank |
|---|---|---|---|
| ОТМЕТКА2 | спорт | 33 | 38.18 |
| | компьютер | 37 | 54.28 |
| | искусство | 30 | 59.38 |
| | Total | 100 | |

**Test Statistics[a,b]**

| | ОТМЕТКА2 |
|---|---|
| Chi-Square | 9.437 |
| df | 2 |
| Asymp. Sig. | .009 |

**a. Kruskal Wallis Test**

**b. Grouping Variable: ХОББИ**

**Рис. 12.17.** Фрагменты окна вывода после выполнения шага 5ж

В первой таблице для каждой группы представлена ее численность и средний ранг. Во второй таблице указано значение критерия $x^2$. число степеней свободы и уровень статистической значимости. Результаты обработки показывают статистически достоверную связь внешкольных увлечений учащихся с успеваемостью в выпускном классе.

# Сравнение К зависимых выборок и критерий Фридмана

*Критерий Фридмана* является непараметрическим аналогом однофакторного дисперсионного анализа для повторных измерений. Он позволяет проверять гипотезы о различии более двух зависимых выборок (повторных измерений) по уровню выраженности изучаемой переменной. Критерий Фридмана может быть более эффективен, чем его метрический аналог однофакторный дисперсионный анализ в случаях повторных измерений изучаемого признака на небольших выборках и при отличии распределений от нормального.

Критерий Фридмана весьма сходен с критерием Краскала—Уоллеса и основан на ранжировании ряда повторных измерений для каждого объекта выборки. Затем вычисляется сумма рангов для каждого из условий (повторных измерений). Если выполняется статистическая гипотеза об отсутствии различий между повторными измерениями, то можно ожидать примерного равенства сумм рангов для этих условий. Чем больше различаются зависимые выборки по изучаемому признаку, тем больше эмпирическое значение вычисляемого значения критерия $\chi^2$, *по* которому определяется *р*-уровень значимости.

Диалоговое окно команды для применения критерия Фридмана представлено на рис. 12.18.

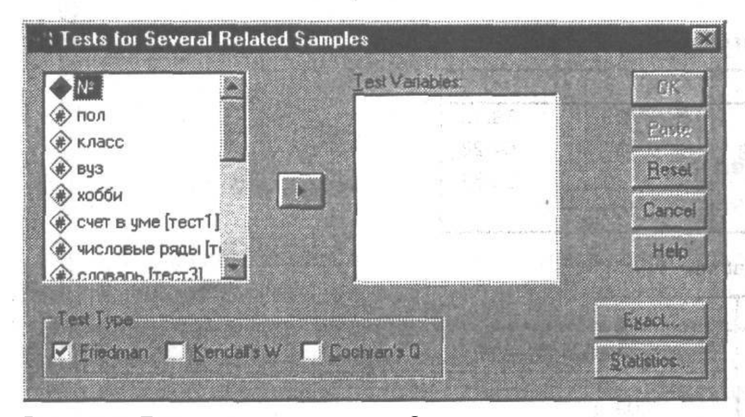



**Рис. 12.18.** Диалоговое окно критерия Фридмана

После завершения шага 3 на экране должно присутствовать окно редактора данных. В этом примере мы сравним результаты тестов тест1, тест2, тест3, тест4 и тест5 для всех учащихся.

**Шаг 5и** В меню Analyze (Анализ) выберите команду Nonparametric Tests ▸ K Related Samples (Непараметрические критерии ▸ К зависимых выборок), чтобы открыть диалоговое окно Test for Several Related Samples (Критерий для нескольких зависимых выборок), показанное на рис. 12.18.

**Шаг 5к** Для применения критерия выполните следующую последовательность действий:

1. Щелкните сначала на переменной тест1, чтобы выделить ее, а затем — на кнопке со стрелкой, чтобы переместить переменную в список Test Variables (Тестовые переменные).
2. Повторите предыдущее действие для переменных тест2, тест3, тест4 и тест5.
3. Щелкните на кнопке OK, чтобы открыть окно вывода.

Результаты работы программы показаны на рис. 12.19.

**Ranks**

| | Mean Rank |
|---|---|
| счет в уме | 2.50 |
| числовые ряды | 2.65 |
| словарь | 3.41 |
| осведомленность | 3.08 |
| кратковременная память | 3.37 |

**Test Statistics[a]**

| | |
|---|---|
| N | 100 |
| Chi-Square | 29.696 |
| df | 4 |
| Asymp. Sig. | .000 |

a. Friedman Test

**Рис. 12.19.** Фрагменты окна вывода после выполнения шага 5з

Средние ранги (Mean Rank) определяются следующим образом: сначала для каждого объекта значения сравниваемых переменных ранжируются. Затем для каждой из сравниваемых переменных вычисляется средний ранг по всем объектам. Определяемый по критерию $\chi^2$ уровень значимости (Asymp. Sig.) равен 0. Он свидетельствует о статистически значимой разности между пятью результатами тестирования. Различаться может любая пара переменных, и без попарного сравнения невозможно выяснить, какие именно пары вносят значимый вклад в статистически достоверный результат.

После выполнения шага 5 программа автоматически активизирует окно вывода. Для просмотра результатов вы при необходимости можете воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

# Печать результатов и выход из программы

Ниже описана типичная процедура печати результатов статистического анализа (или последовательности анализов). После выполнения шага 5 должно быть открыто окно вывода.

**Шаг 6** В окне вывода укажите фрагменты, выводимые на печать (см. раздел «Окно вывода» в главе 2), в меню File (Файл) выберите команду Print (Печать), при необходимости задайте параметры печати и щелкните на кнопке ОК.

Последнее, что необходимо сделать после завершения исследования и печати результатов, — это выйти из программы SPSS.

**Шаг 7** Для выхода из программы в меню File (Файл) выберите команду Exit (Выход).

Иногда после выполнения команды Exit (Выход) на экране могут появляться небольшие диалоговые окна с вопросом о необходимости сохранения сделанных в файлах изменений и кнопками, описывающими возможные варианты ответа. Для завершения работы просто щелкайте на соответствующих кнопках.

# 13 Однофакторный дисперсионный анализ



Однофакторный дисперсионный анализ в SPSS реализуется с помощью команды One-way ANOVA (Однофакторный дисперсионный анализ). Команды подменю General Linear Model (Общие линейные модели), описываемые в главах 14 и 15, также отчасти позволяют проводить подобный анализ, однако их возможности несколько уже, чем у команды One-way ANOVA (Однофакторный дисперсионный анализ).

*Дисперсионный анализ* (Analysis Of Variances, ANOVA) — это процедура сравнения средних значений выборок, на основании которой можно сделать вывод о соотношении средних значений генеральных совокупностей. Ближайшим и более простым аналогом ANOVA является $t$-критерий, применение которого было рассмотрено в главе 11. В отличие от $t$-критерия дисперсионный анализ предназначен для сравнения не двух, а нескольких выборок. Слово «дисперсионный» в названии указывает на то, что в процессе анализа сопоставляются компоненты дисперсии изучаемой переменной. Общая изменчивость переменной раскладывается на две составляющие: межгрупповую (факторную), обусловленную различием групп (средних значений), и внутригрупповую (ошибки), обусловленную случайными (неучтенными) причинами. Чем больше частное от деления межгрупповой и внутригрупповой изменчивости ($F$-отношение) тем больше различаются средние значения сравниваемых выборок и тем выше статистическая значимость этого различия.

Итак, название указывает на то, что вывод о различии средних значений делается на основе анализа компонентов дисперсии. А что означает слово «Однофакторный»? Применяя команду One-way ANOVA (Однофакторный дисперсионный анализ), вы увидите, что в ней можно задать единственную зависимую переменную (при этом она обязательно должна быть количественного, а точнее, метрического типа) и единственную независимую переменную (всегда номинативную, имеющую несколько градаций). Различные модели дисперсионного анализа,

описанные в двух следующих главах, допускают наличие нескольких независимых переменных. Многомерный дисперсионный анализ, с которым вы столкнетесь в главах 15 и 16, позволяет анализировать как множество независимых, так и множество зависимых переменных.

Итак, при однофакторном дисперсионном анализе сравниваются средние значения каждой выборки друг с другом и вычисляется общий уровень значимости различий. Обратите внимание, что вывод по результатам ANOVA касается общего различия всех сравниваемых средних без конкретизации того, какие именно выборки различаются, а какие нет. Для идентификации пар выборок, отличающихся друг от друга средними значениями, используются методы парных сравнений постфактум (post hoc), а для более сложных сопоставлений — метод контрастов (contrasts).

# Пошаговые алгоритмы вычислений

Для проведения однофакторного дисперсионного анализа мы будем использовать файл ex01.sav. В роли зависимой переменной выступит переменная тест1, а независимая переменная класс разделит объекты на 3 выборки, средние значения которых мы будем сравнивать, но сначала необходимо выполнить три подготовительных шага.

**Шаг 1** Создайте новый файл данных или подготовьте существующий. Все необходимые действия описаны в главе 3.

**Шаг 2** Запустите программу SPSS при помощи значка на рабочем столе или команды Пуск ▸ Программы ▸ SPSS for Windows ▸ SPSS 11.5 for Windows (Start ▸ Programs ▸ SPSS for Windows ▸ SPSS 11.5 for Windows) главного меню Windows. В открывшемся после запуска программы диалоговом окне SPSS for Windows щелкните на кнопке Cancel (Отмена).

После выполнения этого шага на экране появится окно редактора данных SPSS.

**Шаг 3** Откройте файл данных, с которым вы намерены работать (в пашем случае это файл ex01.sav). Если он расположен в текущей папке, то выполните следующие действия:

1. Выберите в меню File (Файл) команду Open ▸ Data (Открытие ▸ Данные) или щелкните на кнопке Open File (Открытие файла) панели инструментов.
2. В открывшемся диалоговом окне дважды щелкните на имени ex01.sav или введите его с клавиатуры и щелкните на кнопке OK.

Независимо от того, открыта программа SPSS только что или какие-то процедуры уже выполнялись, в верхней части главного окна должна присутствовать

строка меню (она показана ниже). Пока строка меню присутствует на экране, доступны все команды анализа данных. При этом окно с данными видеть не обязательно.

File Edit View Data Transform Analyze Graphs Utilities Window Help

При работе с таблицами или при редактировании диаграмм некоторые пункты строки меню могут исчезать или меняться. Чтобы вернуться к главному окну и стандартной строке меню, щелкните на кнопке свертывания или восстановления текущего окна.

## Однофакторный дисперсионный анализ

После завершения шага 3 на экране должно присутствовать окно редактора данных со строкой меню.

Шаг 4 В меню Analyze (Анализ) выберите команду Compare Means ▸ One-Way ANOVA (Сравнение средних ▸ Однофакторный дисперсионный анализ). На экране появится диалоговое окно One-Way ANOVA (Однофакторный дисперсионный анализ), показанное на рис. 13.1.

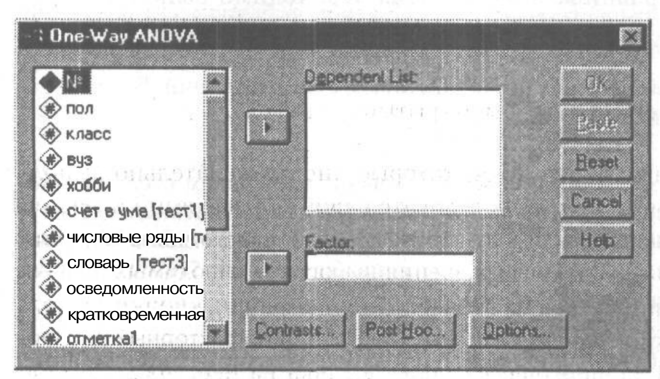



**Рис. 13.1.** Диалоговое окно One-Way ANOVA

Структура диалогового окна One-Way ANOVA (Однофакторный дисперсионный анализ) вполне типична для большинства диалоговых окон SPSS. Слева мы видим список переменных текущего файла данных. В нижней части окна расположены три кнопки: Options (Параметры), Post Hoc (Постфактум) и Contrasts (Контрасты), которые мы будем использовать при обработке. Список Dependent List (Зависимые переменные) предназначен для задания одной или нескольких зависимых переменных (в нашем примере будет использоваться единственная зависимая переменная тест1). Зависимые переменные должны быть метрического типа. Если в списке указано несколько зависимых переменных, то SPSS выполнит ANOVA

для каждой из них. Под списком Dependent List (Зависимые переменные) находится поле Factor (Фактор), в котором нужно указать единственную независимую переменную, имеющую несколько градаций (в нашем случае — хобби). Таким образом мы сравним результаты первого теста («счет в уме») для трех групп учащихся, различающихся внешкольными увлечениями.

Иногда возникает необходимость сравнивать не все выборки, соответствующие градациям фактора (независимой переменной), а лишь часть из них. В этом случае перед проведением анализа необходимо обратиться к команде Select Cases (Выбор объектов) для выбора необходимых для анализа градаций фактора.

На шаге 5 мы будем сравнивать между собой средние значения переменной тест1 для каждой из выборок по уровням переменной хобби.

**Шаг 5** После выполнения шага 4 должно быть открыто диалоговое окно One-Way ANOVA (Однофакторный дисперсионный анализ), показанное на рис. 12.1. При необходимости повторите шаг 4 и выполните следующие действия:

1. Щелкните сначала па переменной тест1, чтобы выделить ее, а затем — па верхней кнопке со стрелкой, чтобы переместить переменную в список Dependent List (Зависимые переменные).
2. Щелкните сначала на переменной хобби, чтобы выделить ее, а затем — на нижней кнопке со стрелкой, чтобы переместить переменную в поле Factor (Фактор).
3. Щелкните на кнопке OK, чтобы открыть окно вывода.

Существует два дополнительных действия, которые иногда желательно выполнять в процессе анализа. Приведенная последовательность инструкций позволяет получить результаты сравнения средних значений выборок, однако ни сами средние значения, пи результаты парного сравнения выборок в выводимых данных отображены не будут. Первая проблема решается при помощи кнопки Options (Параметры). Диалоговое окно One-Way ANOVA: Options (Однофакторный дисперсионный анализ: Параметры), появляющееся после щелчка на ней, представлено на рис. 13.2. Установка флажка Descriptive (Описательные статистики) приведет к включению в выводимые данные всех средних значений, стандартных отклонений, стандартных ошибок, границ доверительных интервалов в 95 %, а также минимумов и максимумов выборок. Флажок Homogeneity of variance test (Критерий однородности дисперсии) позволяет вывести информацию о степени пригодности данных к дисперсионному анализу, а с помощью флажка Means plot (График средних) можно построить диаграмму, на которой будут изображены средние значения для каждой выборки. Группа переключателей Missing Values (Пропущенные значения) позволяет выбрать способ обработки отсутствующих значений (см. раздел «Обработка пропущенных значений» в главе 4).

## Парные сравнения

Нередко нас могут заинтересовать результаты парных сравнений градаций независимой переменной, и для этой цели в диалоговом окне One-Way ANOVA (Однофакторный дисперсионный анализ) предусмотрена специальная кнопка Post Hoc (Постфактум). Слово «постфактум» означает, что эта процедура проводится после установления статистически достоверного результата однофакторного дисперсионного анализа. Если результаты ANOVA оказались статистически недостоверными, применение процедуры парных сравнений некорректно. При щелчке на кнопке Post Hoc (Постфактум) открывается диалоговое окно One-Way ANOVA: Post Hoc Multiple Comparisons (Однофакторныи дисперсионный анализ: Множественные сравнения постфактум), представленное на рис. 13.3. Это диалоговое окно с помощью флажков позволяет задать 14 критериев для выборок с одинаковой дисперсией и 4 критерия для выборок с разной дисперсией. Число предлагаемых критериев настолько велико, что даже руководство пользователя программы SPSS объемом около 3000 страниц оказалось не в состоянии вместить описания их всех.

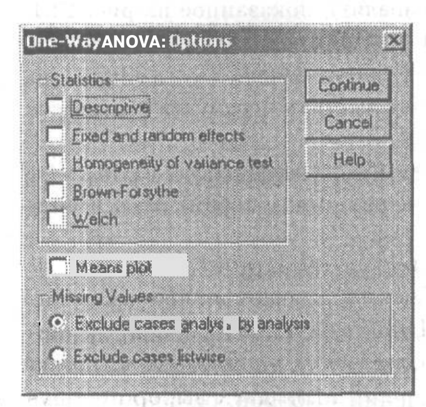



Рис. 13.2. Диалоговое окно One-Way ANOVA: Options

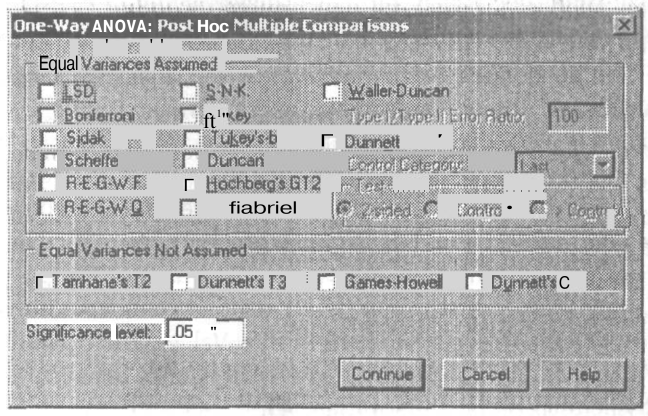



Рис. 13.3. Диалоговое окно One-Way ANOVA: Post Hoc Multiple Comparisons

Большинство из указанных тестов используются очень редко, поэтому ниже приведены описания только нескольких наиболее популярных.

- LSD (Least Significant Difference — Наименьшая значимая разность) — этот тест представляет собой не что иное, как совокупность $t$-критериев для всех возможных пар градаций фактора. Критерий наименьшей значимой разности является одним из самых «либеральных», поскольку наиболее подвержен ошибкам. Например, если независимая переменная имеет 5 уровней, то будет проведено 10 сравнений. При уровнях значимости каждого из сравнений, равных 0,05, существует вероятность почти в 40 % того, что хотя бы 1 из тестов показал значимый результат случайно.
- Bonferroni (Бонферрони) — критерий Бонферрони схож с критерием наименьшей значимой разности, однако лишен недостатка, связанного с повторными проверками: в нем уровень значимости делится на число сравнений. Таким образом, критерий Бонферрони является более «консервативным».
- Scheffe (Шеффе) — критерий Шеффе еще более «консервативен», чем критерий Бонферрони, использующий $F$-критерий вместо $t$-критерия.
- Tukey (Тьюки) — критерий Тьюки использует статистику Стьюдента (Student) для определения различий между группами. Этот критерий часто применяется в случаях, когда исследуемый фактор имеет большое количество уровней.

Самыми консервативными из предложенных являются критерии Шеффе и Бонферрони. Часто используется также критерий Тьюки, называемый еще критерием подлинной значимости (Honestly Significant Difference, HSD). HSD представляет собой наименьшую величину разности средних значений выборок, которую можно считать значимой. Например, если HSD - 2,5, а для двух выборок получены величины средних значений 3,7 и 6,3, то разность между ними, равная 2,6, согласно Тьюки, является значимой, поскольку она превышает величину HSD. При использовании критерия Тьюки программа SPSS также включает в вывод дополнительную статистическую информацию.

Необходимо отметить, что все описанные выше критерии (а на самом деле большинство критериев парных сравнений) применяются в предположении, что дисперсии всех ячеек равны. Исключение составляют 4 критерия, флажки для которых выделены в отдельную группу в нижней части диалогового окна One-Way ANOVA: Post Hoc Multiple Comparisons (Однофакторный дисперсионный анализ: Множественные сравнения постфактум), — они применяются в случаях, когда дисперсии ячеек разные.

В следующем примере мы проведем однофакторный дисперсионный анализ, а в выводимые результаты включим описательные статистики и критерий однородности дисперсии. Для парных сравнений воспользуемся критерием Шеффе.

Как и в предыдущем примере, зависимой переменной будет переменная тест1, а независимой — переменная хобби.

**Шаг 5а** После выполнения шага 4 должно быть открыто диалоговое окно One-Way ANOVA (Однофакторный дисперсионный анализ), показанное на рис. 13.1. Если вы уже успели поработать с этим окном, очистите его щелчком на кнопке Reset (Сброс) и выполните следующие действия:

1. Щелкните сначала на переменной тест1, чтобы выделить ее, а затем — на Верхней кнопке со стрелкой, чтобы переместить переменную в список Dependent List (Зависимые переменные).
2. Щелкните сначала на переменной хобби, чтобы выделить ее, а затем — на нижней кнопке со стрелкой, чтобы переместить переменную в поле Factor (Фактор).
3. Щелкните на кнопке Options (Параметры), чтобы открыть диалоговое окно One-WayANOVA: Options (Однофакторный дисперсионный анализ: Параметры), показанное на рис. 13.2.
4. Установите флажки Descriptive (Описательные статистики) и Homogeinity ofvariance test (Критерий однородности дисперсии), а затем щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно One-Way ANOVA (Однофакторный дисперсионный анализ).
5. Щелкните на кнопке Post Hoc (Постфактум), чтобы открыть диалоговое окно One-Way ANOVA: Post Hoc Multiple Comparisons (Однофакторный дисперсионный анализ: Множественные сравнения постфактум), показанное на рис. 13.3.
6. Установите флажок Sheffe (Шеффе) и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно One-Way ANOVA (Однофакторный дисперсионный анализ).
7. Щелкните на кнопке OK, чтобы открыть окно вывода.

## Контрасты

Последняя из трех кнопок в окне One-Way ANOVA (Однофакторный дисперсионный анализ) имеет название Contrasts (Контрасты). Она предназначения для вызова диалогового окна One-Way ANOVA: Contrasts (Однофакторный дисперсионный анализ: Контрасты), представленного на рис. 13.4. Это окно позволяет осуществлять различные сравнения выборок по градациям независимой переменной. Так, вы можете сравнивать одну градацию с другой, одну градацию со всеми остальными или разбить все градации на 2 группы и затем сравнить их между собой. Сравнение сводится к применению модифицированного варианта $t$-критерия. Отметим, что в отличие от процедуры парного сравнения постфактум применение метода контрастов не требует предварительного получения статистически достоверного результата анализа ANOVA.

В нашем примере независимая переменная хобби имеет 3 градации: 1 — спорт, 2 — компьютер, 3 — искусство. Для задания контраста предназначены поле и список Coefficients (Коэффициенты). Заполнение списка происходит следующим образом. Каждой градации фактора вы должны сопоставить число, определяющее его роль в контрасте: отрицательное число соответствует одной группе, положительное число — другой группе, а ноль означает, что градация в сравнениях не задействована. При этом абсолютные величины коэффициентов не важны, но последние должны вводиться в порядке следования градаций и в сумме давать нулевое значение. Например, если вам необходимо сравнить увлекающихся спортом с остальными учащимися, вы можете закодировать градации последовательностью –2, 1 и 1, а если вы хотите сравнить увлекающихся спортом только с теми, кто увлекается искусством, то необходимо задать последовательность 1, 0 и -1. Не забывайте, что сумма всех коэффициентов обязательно должна равняться 0.

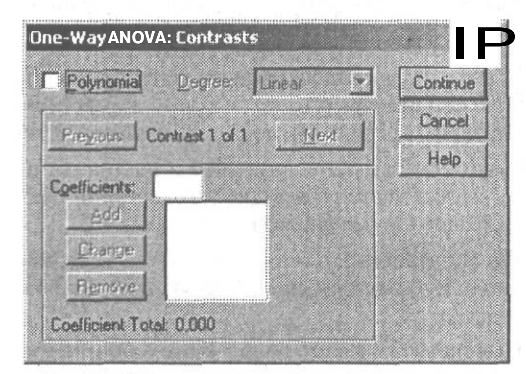



**Рис. 13.4.** Диалоговое окно One-Way ANOVA: Contrasts

Помещение чисел в список осуществляется путем ввода значения в поле справа от названия списка и щелчка на кнопке Add (Добавление). Если нужно создать несколько «контрастов», то есть разбиений на группы, то щелкните на кнопке Next (Следующий) справа от метки Contrast 1 of 1 (Контраст 1 из 1) и повторите описанную процедуру.

Следующий пример иллюстрирует применение контрастов для сравнения учащихся, увлекающихся компьютером, с теми, кто имеет другие увлечения, а также для сравнения увлекающихся спортом с остальными учащимися.

После выполнения шага 5 программа автоматически активизирует окно вывода. Для просмотра результатов вы при необходимости можете воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

Шаг 5б После выполнения шага 4 должно быть открыто диалоговое окно One-Way ANOVA (Однофакторный дисперсионный анализ), показанное на рис. 13.1. Если вы уже успели поработать с этим окном, очистите его щелчком на кнопке Reset (Сброс) и выполните следующие действия:

1. Щелкните сначала на переменной тест1, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в список Dependent List (Зависимые переменные).
2. Щелкните сначала на переменной хобби, чтобы выделить ее, а затем — на нижней кнопке со стрелкой, чтобы переместить переменную в поле Factor (Фактор).
3. Щелкните на кнопке Options (Параметры), чтобы открыть диалоговое окно One-Way ANOVA: Options (Однофакторный дисперсионный анализ: Параметры), показанное на рис. 13.2.
4. Установите флажки Descriptive (Описательные статистики) и Homogeinity of variance test (Критерий однородности дисперсии), а затем щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно One-Way ANOVA (Однофакторный дисперсионный анализ).
5. Щелкните на кнопке Post Hoc (Постфактум), чтобы открыть диалоговое окно One-Way ANOVA: Post Hoc Multiple Comparisons (Однофакторный дисперсионный анализ: Множественные сравнения постфактум), показанное на рис. 13.3.
6. Установите флажок LSD (Наименьшая значимая разность) и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно One-Way ANOVA (Однофакторный дисперсионный анализ).
7. Щелкните на кнопке Contrasts (Контрасты), чтобы открыть диалоговое окно One-Way ANOVA: Contrasts (Однофакторный дисперсионный анализ: Контрасты), представленное на рис. 13.4.
8. Нажмите клавишу Tab, чтобы перевести фокус ввода в поле Coefficients (Коэффициенты), введите число 1 и щелкните на кнопке Add (Добавление), задав первое число в списке.
9. Повторите предыдущее действие сначала для чисел -2 и 1, затем, щелкнув на кнопке Next (Следующий), — для чисел -2, 1, 1, после чего щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно One-Way ANOVA (Однофакторный дисперсионный анализ).
10. Щелкните на кнопке OK, чтобы открыть окно вывода.

# Печать результатов и выход из программы

Далее описана типичная процедура печати результатов статистического анализа (или последовательности анализов). После выполнения шага 5 должно быть открыто окно вывода.



Шаг 6 В окис вывода укажите фрагменты, выводимые на печать (см. раздел «Окно вывода» в главе 2), в меню File (Файл) выберите команду Print (Печать), при необходимости задайте параметры печати и щелкните на кнопке OK.

Последнее, что необходимо сделать после завершения исследования и печати результатов, — это выйти из программы SPSS.

**Шаг 7** Для выхода из программы в меню File (Файл) выберите команду Exit (Выход).

Иногда после выполнения команды Exit (Выход) на экране могут появляться небольшие диалоговые окна с вопросом о необходимости сохранения сделанных в файлах изменений и кнопками, описывающими возможные варианты ответа. Для завершения работы просто щелкайте на соответствующих кнопках.

# Представление результатов

В этом разделе мы приводим некоторые результаты процедур, описываемых в этой главе.

## Однофакторный дисперсионный анализ

На рис. 13.5 приведен фрагмент выводимых данных, относящийся к однофакторному дисперсионному анализу и сгенерированный программой после выполнения шага 5б.

**ANOVA**

счет в уме

| | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Between Groups | 82.969 | 2 | 41.484 | 6.556 | .002 |
| Within Groups | 613.781 | 97 | 6.328 | | |
| Total | 696.750 | 99 | | | |

**Рис. 13.5.** Фрагмент окна вывода после выполнения шага 5б, относящийся к однофакторному дисперсионному анализу

Самым важным значением в этой таблице является уровень значимости $p = 0{,}002$. Он указывает на то, что разность между средними значениями переменной тест1 для трех групп статистически достоверна. Далее дана трактовка терминов, используемых программой в окне вывода.

- Значение в столбце Sum of Squares (Сумма квадратов) строки Between Croups (Между группами) представляет собой сумму квадратов разностей между общим средним значением и средними значениями каждой группы, умноженными на весовые коэффициенты, равные числу объектов в группе, а строки Within Croups (Между группами) — сумму квадратов разностей среднего значения каждой группы и каждого значения этой группы.
- Значение в столбце df (Число степеней свободы) строки Between Groups (Между группами) представляет собой межгрупповое число степеней свободы, равное числу групп, уменьшенному на 1, в строки Within Croups (Между группами) -

внутригрупповое число степеней свободы, равное разности между числом объектов и числом групп, уменьшенной на 1.

- Mean Square (Средний квадрат) — отношение суммы квадратов к числу степеней свободы.
- F (*F*-критерий) — отношение среднего квадрата между группами к среднему квадрату внутри группы.
- Sig. (Значимость) — вероятность того, что наблюдаемые различия случайны. Величина значимости $p = 0{,}002$ свидетельствует о статистически достоверных различиях.

## Описательные статистики

На рис. 13.6 приведен фрагмент выводимых данных, сгенерированный программой после выполнения шага 5б и относящийся к описательным статистикам.

**Descrlptives**

счет в уме

| | N | Mean | Std. Deviation | Std. Error | 95% Confidence Interval for Mean | | Minimum | Maximum |
|---|---|---|---|---|---|---|---|---|
| | | | | | у | И | | |
| Спорт | 33 | 9.67 | 2.582 | .449 | 8.75 | 10.58 | 4 | 14 |
| Компьютер | 37 | 11.43 | 2.102 | .346 | 10.73 | 12.13 | 7 | 17 |
| Искусство | 30 | 9.43 | 2.885 | .527 | 8.36 | 10.51 | 5 | 15 |
| Total | 100 | 10.25 | 2.653 | .265 | 9.72 | 10.78 | 4 | 17 |

**Рис. 13.6.** Фрагмент окна вывода после выполнения шага 5б, относящийся к описательным статистикам

# Терминология, используемая при выводе

Трактовка терминов, используемых программой в окне вывода, дана далее.

- N (Число объектов) — число объектов для каждой из градаций переменной хобби.
- Mean (Среднее) — среднее значение переменной тест1 для каждой группы.
- Std. Deviation (Стандартное отклонение) — мера разброса значений распределения относительно среднего.

- Std. Error (Стандартная ошибка) — отношение стандартного отклонения к квадратному корню из числа объектов.
- 95% Confidence Interval for Mean (Доверительный интервал для среднего значения в 95 %) — при большом числе выборок из генеральной совокупности 95 % средних значений этих выборок попадут в интервал, определяемый указанными в таблице границами.
- Minimum (Минимум) — наименьшее из наблюдаемых значений для группы.
- Maximum (Максимум) — наибольшее из наблюдаемых значений для группы.

## Парные сравнения постфактум

На рис. 13.7 приведен фрагмент выводимых данных, сгенерированный программой после выполнения шага 5а и относящийся к парным сравнениям.

**Multiple Comparisons**

Dependent Variable: счет в уме
Scheffe

| (I) ХОББИ | (J) ХОББИ | Seac Difference (I-J) | Std. Error | Sig. | 95% Confidence Interval | |
|---|---|---|---|---|---|---|
| | | | | | Lower Bound | Upper Bound |
| Спорт | Компьютер | -1.77* | .602 | .016 | -3.26 | -.27 |
| | Искусство | .23 | .635 | .935 | -1.34 | 1.81 |
| Компьютер | Спорт | 1.77* | .602 | .016 | .27 | 3.26 |
| | Искусство | 2.00* | .618 | .007 | .46 | 3.54 |
| Искусство | Спорт | -.23 | .635 | .935 | -1.81 | 1.34 |
| | Компьютер | -2.00* | .618 | .007 | -3.54 | -.46 |

*• The mean difference is significant at the .05 level.

**Рис. 13.7.** Фрагмент окна вывода после выполнения шага 5а, относящийся к парным сравнениям

Средние значения переменной тест1 для каждой их трех выборок были перечислены в предыдущей таблице (см. рис. 13.6), здесь же даны разности между этими значениями. Знаком звездочки помечены те пары выборок, для которых разность средних значений статистически достоверна, то есть со значением уровня значимости 0,05 и меньше. Из полученных данных можно сделать вывод, что результаты

теста 1 для тех, кто увлекается компьютером, статистически значимо выше, чем для тех, кто увлекается спортом и искусством. Те же, кто увлекаются спортом и искусством, по результатам теста 1 статистически достоверно не различаются.

## Критерий Левина

На рис. 13.8 приведен фрагмент выводимых результатов, сгенерированный программой после выполнения шага 5б и относящийся к критерию Левина.

**Test of Homogeneity of Variances**

счет в уме

| Levene Statistic | df1 | df2 | Sig. |
|---|---|---|---|
| 1.858 | 2 | 97 | .161 |

**Рис. 13.8.** Фрагмент окна вывода после выполнения шага 5б, относящийся к критерию Левина

Критерий однородности дисперсии Левина со значимостью 0,161 показал, что дисперсии для каждой из групп статистически достоверно не различаются. Следовательно, результаты ANOVA могут быть признаны корректными. Если бы результат применения критерия Левина оказался статистически достоверным, то это послужило бы основанием для сомнения в корректности применения анализа ANOVA.

## Контрасты

На рис. 13.9 приведены фрагменты выводимых результатов, сгенерированные программой после выполнения шага 5б и относящиеся к применению метода контрастов.

Первая из таблиц, приведенных на рисунке, содержит коэффициенты, введенные при группировании уровней. Для каждого контраста было применено по два $t$-критерия, один из которых проводился с допущением о равных дисперсиях, а другой — с допущением о неравных дисперсиях. Первый контраст оказался статистически достоверным: те, кто увлекается компьютером, имеют статистически достоверно более высокий результат по переменной тест1. Второй контраст не достигает статистически достоверного уровня: учащиеся, увлекающиеся спортом, статистически достоверно не отличаются от других учащихся по переменной тест1. Ниже дана трактовка терминов, используемых программой в окне вывода.

- Value of Contrast (Величина контраста) — это значение не представляет интереса для исследователя, поскольку является всего лишь весовым коэффициентом.
- Std. Error (Стандартная ошибка) — отношение стандартного отклонения контраста к квадратному корню из числа объектов.
- t ($t$-критерий) — отношение величины контраста к стандартной ошибке.

**Contrast Coefficients**

| Contrast | ХОББИ | | |
|---|---|---|---|
| | Спорт | Компьютер | Искусство |
| 1 | 1 | -2 | 1 |
| 2 | -2 | 1 | 1 |

Contrast Tests

| | | Contrast | Value of Contrast | Std. Error | t | df | Sig. (2-tailed) |
|---|---|---|---|---|---|---|---|
| счет в уме | Assume equal variances | 1 | -3.76 | 1.042 | -3.611 | 97 | .000 |
| | | 2 | 1.53 | 1.072 | 1.430 | 97 | .156 |
| | Does not assume equal variances | 1 | -3.76 | .978 | -3.848 | 89.220 | .000 |
| | | 2 | 1.53 | 1.098 | 1.396 | 61.897 | .168 |

**Рис. 13.9.** Фрагменты окна вывода после выполнения шага 5б, относящиеся к применению метода контрастов

- df (Число степеней свободы) — величина, для равных дисперсий равная разности числа объектов и числа групп, уменьшенной па единицу. В случае неравных дисперсий число степеней свободы вычисляется по более сложной формуле и может быть дробным.
- Sig. (2-tailed) (Двусторонняя значимость) — вероятность того, что отличие контраста от нуля является случайным.

# 14 Многофакторный дисперсионный анализ



В этой главе речь пойдет о дисперсионном анализе (ANOVA) с двумя и более факторами. В нем участвует единственная зависимая переменная, которая, как и при однофакторном дисперсионном анализе, должна быть метрической. Главное отличие многофакторного анализа от однофакторного заключается в том, что в нем участвуют не одна, а несколько независимых переменных, каждая из которых представлена в номинативной шкале (имеет несколько градаций, или уровней). Многофакторный дисперсионный анализ реализуется в SPSS с помощью команды General Linear Model ▸ Univariate (Общие линейные модели ▸ Одномерный анализ).

Всего дисперсионному анализу посвящены главы 13-16. Мы настоятельно рекомендуем читать эти главы по порядку, поскольку в них в развитии описывается одна и та же статистическая концепция. Представление об однофакторном анализе является основой для понимания дисперсионного анализа с двумя и более факторами, который, в свою очередь, позволяет перейти к более сложным моделям дисперсионного анализа. Отметим, что дисперсионный анализ является очень сложным разделом статистики, поэтому в этой книге мы приводим лишь самые необходимые сведения, оставляя для самостоятельного изучения все специфические вопросы.

Программа SPSS предназначена для того, чтобы сделать работу исследователя как можно более простой и комфортной. Чтобы выполнить дисперсионный анализ с одним, двумя и даже семью факторами, достаточно лишь нескольких щелчков на нужных кнопках. Компьютер выполнит за вас всю вычислительную работу, и результат вы получите в наглядной и удобной форме, но, к сожалению, нередко совершенно неясный по своему содержанию. Простота и удобство вычислительного процесса подчас приводят к тому, что исследователи «расслабляются» и за-

бывают простую истину: успех исследования определяется его продуманностью. Если однофакторный дисперсионный анализ не представляет сложности для понимания, то анализ с двумя и более факторами требует очень внимательного отношения и определенных навыков в интерпретации результатов. Так, для интерпретации результатов трехфакторного дисперсионного анализа, как правило, необходимы немалый опыт и умение свободно обращаться со статистическими величинами. Что же касается четырехфакторного анализа, то его интерпретация чрезвычайно сложна даже для специалиста высокой квалификации.

Многофакторный дисперсионный анализ отличается от однофакторного появлением одной общей проблемы — проблемы взаимодействия факторов. Решение этой проблемы принципиально не зависит от числа факторов. С возрастанием числа факторов существенно нарастает лишь сложность интерпретации взаимодействий. Поэтому мы подробно рассмотрим только самый распространенный вариант дисперсионного анализа — анализ с двумя факторами, предполагая, что полученные умения вы сможете распространить и на анализ с большим числом факторов.

В этой главе мы выполним двухфакторный дисперсионный анализ и анализ с учетом влияния ковариаты. Мы также воспользуемся графическими средствами интерпретации средних значений и, как всегда, в конце главы проведем разбор полученных результатов.

# Дисперсионный анализ с двумя факторами

Как было сказано в предыдущей главе, дисперсионный анализ определяет статистическую достоверность различия между выборками путем сравнения их средних значений. Чтобы получить представление о двухфакторном анализе, сначала немного вернемся назад и обобщим наши знания об однофакторном дисперсионном анализе. Мы сравнивали три класса (переменная класс) по уровню выраженности переменной тест1. «Однофакторность» анализа заключалась в том, что деление на группы производилось по градациям одной независимой переменной (класс). Для однофакторного дисперсионного анализа в SPSS существует специальная упрощенная команда One-Way ANOVA (Однофакторный дисперсионный анализ). Команды подменю General Linear Model (Общие линейные модели) позволяют выполнять однофакторный, двухфакторный, трехфакторный и т. д. анализ, однако обращение с ними несколько сложнее.

В этой главе мы будем использовать файл ex01.sav. В качестве зависимой переменной мы возьмем переменную отметка2, а роль независимых будут играть переменные пол и хобби. Мы попытаемся определить степень влияния переменных пол, хобби и их взаимодействия полххобби на распределение значений переменной отметка2. Такая схема анализа может быть обозначена как

ANOVA 2 x 3(полxхобби). Исследование позволит получить ответы на перечисленные ниже вопросы.

- Существует ли главный эффект фактора пол, то есть существует ли значимое различие оценок для юношей и девушек, и какова степень этого различия?
- Существует ли главный эффект фактора хобби, то есть существует ли значимое различие оценок между тремя группами, и какова степень этого различия?
- Существует ли взаимодействие переменных пол и хобби, то есть зависит ли влияние одной из этих переменных от градаций другой?

Таким образом, двухфакторный дисперсионный анализ позволяет проверить три гипотезы: две о главном эффекте и одну о взаимодействии факторов. Ответы на два первых вопроса можно было бы получить, применив дважды однофакторный дисперсионный анализ. Специфика многофакторного анализа проявляется в содержании третьего вопроса, который касается взаимодействия факторов. Взаимодействие двух факторов означает, что влияние одного из них проявляется по-разному на разных уровнях другого фактора. Весьма полезным для интерпретации взаимодействий является построение диаграмм средних значений для каждой ячейки таблицы сопряженности независимых переменных. Мы займемся построением диаграмм в конце этой главы. На диаграммах будут представлены все значения вне зависимости от степени их значимости, что позволит нам более ясно представлять, почему воздействие одних переменных оказывается значимым, а других — нет. Более детально проблема взаимодействия будет рассмотрена на примере при обсуждении результатов.

SPSS может включить в вывод средние значения всех выборок, соответствующих всем возможным сочетаниям градаций факторов (в данном случае $2 \times 3 = 6$), а также вычислить *F*-величины и соответствующие *p*-уровни. По этим характеристикам мы сможем судить о степени влияния каждой из независимых переменных на распределение зависимой переменной и о взаимодействии независимых переменных.

# Дисперсионный анализ с тремя и более факторами

Как уже отмечалось, проведение дисперсионного анализа с тремя и более факторами принципиально не отличается от двухфакторного анализа. При количестве факторов более двух возрастает лишь сложность интерпретации взаимодействий факторов. Эта сложность обусловлена появлением большого числа взаимодействий, некоторые из которых с трудом поддаются интерпретации. Рассмотрим эту проблему на примере.

Предположим, изучается влияние на переменную отметка2 трех факторов: пол, хобби и класс. Такая схема анализа может быть обозначена как ANOVA $2 \times 3 \times 3$

(пол х хобби х класс). Применение трехфакторного дисперсионного анализа позволило бы получить ответы на следующие вопросы:

- Существует ли главный эффект фактора пол, то есть существует ли значимое различие оценок для юношей и девушек, и какова степень этого различия?
- Существует ли главный эффект фактора хобби, то есть существует ли значимое различие оценок между тремя группами, и какова степень этого различия?
- Существует ли главный эффект фактора класс, то есть существует ли значимое различие оценок между классами, и какова степень этого различия?
- Существует ли взаимодействие переменных пол и хобби, то есть зависит ли влияние одной из этих переменных от градаций другой?
- Существует ли взаимодействие переменных пол и класс, то есть зависит ли влияние одной из этих переменных от градаций другой?
- Существует ли взаимодействие переменных класс и хобби, то есть зависит ли влияние одной из этих переменных от градаций другой?
- Существует ли взаимодействие переменных пол, хобби и класс, то есть зависит ли взаимодействие двух из этих переменных от градаций третьей?

Таким образом, трехфакторный дисперсионный анализ предполагает проверку уже семи гипотез. Из них три гипотезы касаются взаимодействия первого порядка (двух факторов) и одна — взаимодействия второго порядка (трех факторов). Если добавить еще один, четвертый фактор, то появится взаимодействие третьего порядка, а число проверяемых гипотез возрастет до 15. Для пяти факторов число проверяемых гипотез составит уже 31! Общее число проверяемых гипотез равно $2P$ - 1, где Р — число факторов.

При увеличении числа факторов быстро возрастает не только количество проверяемых гипотез, но и сложность интерпретации взаимодействий. Если, как мы увидим далее, интерпретация взаимодействия первого порядка (двух факторов) обычно не составляет труда, то взаимодействие второго порядка обязательно требует построения графиков средних значений, а интерпретация взаимодействия третьего порядка и выше вряд ли вообще возможна.

# Влияние ковариат

Ковариаты используются для исключения влияния количественной переменной на зависимую переменную. Ковариату проще всего представить как переменную, значительно коррелирующую с зависимой переменной и позволяющую уменьшить дисперсию последней. Включение в анализ ковариаты по смыслу означает исключение ее влияния на зависимую переменную. За счет этого дисперсия последней уменьшается, что позволяет сделать более очевидным влияние факторов.

В нашем исследовании в качестве ковариаты будет использоваться переменная тест_ср. Эта переменная, как суммарный показатель интеллектуальных

способностей, в существенной степени коррелирует с успеваемостью (отметка2). Если мы хотим проследить влияние факторов пол и хобби на зависимую переменную отметка2 без учета влияния фактора тест_ср, то последний необходимо включить в анализ в качестве ковариаты. Появление ковариаты не влияет на описательные статистики, однако может изменить сумму квадратов (как правило, в меньшую сторону) и величину $F$-критерия (как в меньшую, так и, возможно, в большую сторону).

# Пошаговые алгоритмы вычислений

Для проведения многофакторного дисперсионного анализа сначала необходимо выполнить три подготовительных шага.

Шаг 1 Создайте новый файл данных или подготовьте существующий. Все необходимые действия описаны в главе 3.

Шаг 2 Запустите программу SPSS при помощи значка на рабочем столе или команды Пуск > Программы ▸ SPSS for Windows ▸ SPSS 11.5 for Windows (Start ▸ Programs > SPSS for Windows ▸ SPSS 11.5 for Windows) главного меню Windows. В открывшемся после запуска программы диалоговом окне SPSS for Windows щелкните на кнопке Cancel (Отмена).

После выполнения этого шага на экране появится окно редактора данных SPSS.

Шаг 3 Откройте файл данных, с которым вы намерены работать (в нашем случае это файл ex01.sav). Если он расположен в текущей папке, то выполните следующие действия:

1. Выберите в меню File (Файл) команду Open ▸ Data (Открытие ▸ Данные) или щелкните на кнопке Open File (Открытие файла) панели инструментов.
2. В открывшемся диалоговом окне дважды щелкните па имени ex01.sav или введите его с клавиатуры и щелкните на кнопке OK.

Независимо от того, открыта программа SPSS только что или какие-то процедуры уже выполнялись, в верхней части главного окна должна присутствовать строка меню (она показана ниже). Пока строка меню присутствует на экране, доступны все команды анализа данных. При этом окно с данными видеть не обязательно.

File Edit View Data Transform Analyze Graphs Utilities Window Help

При работе с таблицами результатов или при редактировании диаграмм некоторые пункты строки меню могут исчезать или меняться. Чтобы вернуться к главному окну и стандартной строке меню, щелкните на кнопке свертывания или восстановления текущего окна.

После завершения шага 3 на экране должно присутствовать окно редактора данных со строкой меню.

**Шаг 4** В меню Analyze (Анализ) выберите команду General Linear Model ▸ Univariate (Общие линейные модели ▸ Одномерный анализ). На экране появится диалоговое окно Univariate (Одномерный анализ), показанное на рис. 14.1.

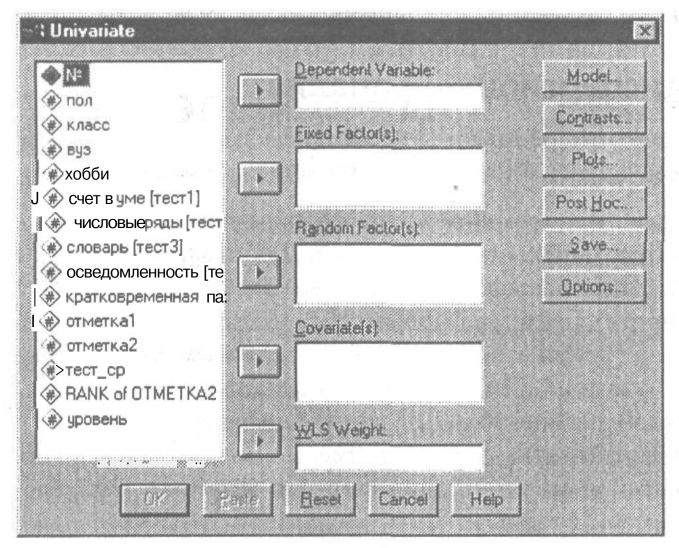



**Рис. 14.1.** Диалоговое окно Univariate

Как вы, вероятно, обратили внимание, в меню General Linear Model (Общие линейные модели), помимо команды Univariate (Одномерный анализ), находятся еще три команды: Multivariate (Многомерный анализ), Repeated Measures (Повторные измерения) и Variance Components (Компоненты дисперсии). Команда Multivariate (Многомерный анализ) используется для многомерного дисперсионного и ковариационного анализов и рассмотрена в главе 15. Описание команды Repeated Measures (Повторные измерения) приведено в главе 16; она может применяться как для однофакторного, так и для многофакторного анализов. Команда Variance Components (Компоненты дисперсии) обладает несколько большей гибкостью по сравнению с предыдущими, однако ее сложность не позволяет рассказать о ней в рамках данной книги.

В левой части диалогового окна Univariate (Одномерный анализ) расположен список всех доступных переменных файла данных, а справа находятся 5 полей, с помощью которых задаются основные параметры дисперсионного анализа. Поле Dependent Variable (Зависимая переменная) предназначено для указания единственной зависимой переменной анализа, поле Fixed factor(s) (Постоянные факторы) — для имен независимых переменных, или факторов, поле Covariate(s) (Ковариаты) — для указания имен ковариат. Кнопки Contrasts (Контрасты) и Post Hoc (Постфактум) действуют практически идентично своим аналогам из диалогового окна One-Way ANOVA (Однофакторный дисперсионный анализ), поэтому при необходимости вернитесь к главе 13.

Особое значение имеет кнопка Plots (Диаграммы), позволяющая строить графики средних значений. Для построения графиков в программе SPSS предусмотрено отдельное меню Graphs (Графики), но в связи с тем, что графики средних значений совершенно необходимы для интерпретации взаимодействий факторов, в основные диалоговые окна команд подменю General Linear Model (Общие линейные модели) включена кнопка Plots (Диаграммы), позволяющая строить графики средних, не выходя из диалогового окна настройки параметров анализа.

## Двухфакторный дисперсионный анализ

После задания зависимой переменной отметка2 необходимо задать независимые переменные, по очереди помещая их в поле Fixed factor(s) (Постоянные факторы).

Когда имена всех переменных, участвующих в анализе, определены, вы можете перейти к заданию параметров выполняемых действий. Для этого щелкните на кнопке Options (Параметры). На экране появится диалоговое окно Univariate: Options (Одномерный анализ: Параметры), представленное на рис. 14.2. Нас будет интересовать группа флажков Display (Отображать), позволяющая задать список включаемых в вывод величии. Так, флажок Descriptive statistics (Описательные статистики) позволяет вывести для каждой ячейки таблицы сопряженности градаций факторов средние значения, стандартные отклонения и размеры выборок. Нередко используется флажок Estimates of effect size (Оценка величины эффекта), включающий в выводимые данные величину эффекта ($\eta^2$), предназначенную для оценки влияния каждой из независимых переменных, а также их взаимодействий.

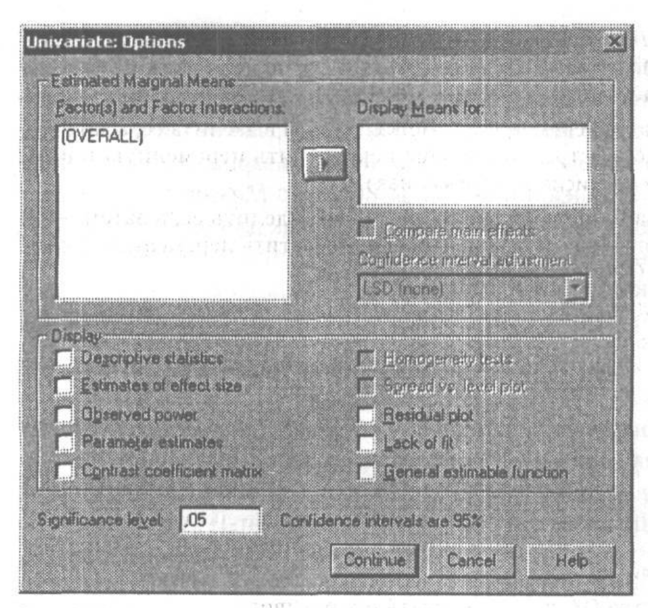



Рис. 14.2. Диалоговое окно Univariate: Options

Следующий пример представляет собой практическую реализацию описанных действий.

**Шаг 5** После выполнения шага 4 должно быть открыто диалоговое окно Univariate (Одномерный анализ), показанное на рис. **14.1.**

1. Щелкните сначала на переменной отметка2, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в поле Dependent Variable (Зависимая переменная).
2. Щелкните сначала на переменной пол, чтобы выделить ее, а затем — на второй сверху кнопке со стрелкой, чтобы переместить переменную в список Fixed Factor(s) (Постоянные факторы).
3. Повторите предыдущее действие для переменной хобби.
4. Щелкните на кнопке Options (Параметры), чтобы открыть диалоговое окно Univariate: Options (Одномерный анализ: Параметры).
5. Установите флажки Descriptive statistics (Описательные статистики) и Estimates of effect size (Оценка величины эффекта), а затем щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Univariate (Одномерный анализ).
6. Щелкните на кнопке OK, чтобы открыть окно вывода.

## Влияние ковариаты

В следующем пошаговом алгоритме мы реализуем двухфакторный дисперсионный анализ с учетом ковариаты. Все переменные останутся теми же, но дополнительно в качестве ковариаты включим переменную тест_ср. Это позволит сравнить результаты дисперсионного анализа с учетом и без учета влияния ковариаты.

**Шаг 5а** После выполнения шага 4 должно быть открыто диалоговое окно Univariate (Одномерный анализ), показанное на рис. 14.1. Если вы уже успели поработать с этим окном, щелкните на кнопке Reset (Сброс).

1. Щелкните сначала на переменной отметка2, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в поле Dependent Variable (Зависимая переменная).
2. Щелкните сначала на переменной пол, чтобы выделить ее, а затем — на второй сверху кнопке со стрелкой, чтобы переместить переменную в список Fixed Factor(s) (Постоянные факторы).
3. Повторите предыдущее действие для переменной хобби.
4. Щелкните сначала на переменной тест_ср, чтобы выделить ее, а затем — на второй снизу кнопке со стрелкой, чтобы переместить переменную в список Covariates (Ковариаты).
5. Щелкните на кнопке Options (Параметры), чтобы открыть диалоговое окно Univariate: Options (Одномерный анализ: Параметры).
6. Установите флажки Descriptive statistics (Описательные статистики) и Estimates of effect size (Оценка величины эффекта), а затем щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Univariate (Одномерный анализ).
7. Щелкните на кнопке OK, чтобы открыть окно вывода.

Результаты выполнения шага 5а будут отличаться от результатов выполнения шага 5 только учетом ковариаты тест_ср.

## Графические средства интерпретации взаимодействий

В большинстве случаев для интерпретации взаимодействия факторов необходимо построение графиков средних значений. Для этого служит кнопка Graphs (Графики), открывающая диалоговое окно Univariate: Profile Plots (Одномерный анализ: Диаграммы профилей), показанное па рис. 14.3. В этом окне можно задать параметры сразу нескольких графиков средних значений. Каждый график представляет собой ломаную линию (профиль), соединяющую точки — средние значения для групп. Горизонтальной оси соответствуют градации одного из факторов, а величина средних значений откладывается по вертикальной оси.

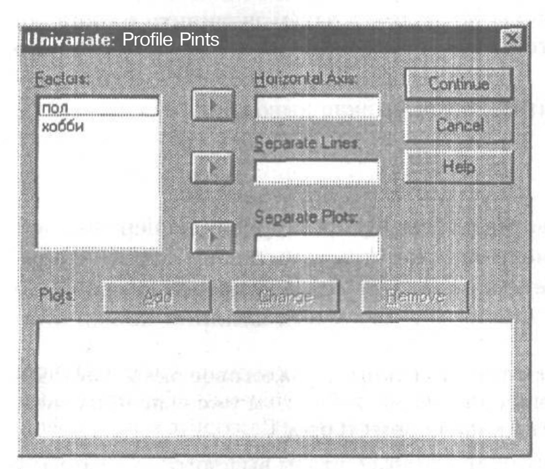



**Рис. 14.3.** Диалоговое окно Univariate: Profile Plots

Вид графиков задается тремя параметрами. В поле Horisontal Axis (Горизонтальная ось) вводится имя фактора, градациям которого будет соответствовать горизонтальная ось графика или графиков. В ноле Separate Lines (Отдельные линии) указывается фактор, градациям которого будут соответствовать разные линии на графике. Поле Separate Plots (Отдельные диаграммы) служит для задания имени фактора, градациям которого будут соответствовать разные графики. Щелчок на кнопке Add (Добавление) подтверждает правильность введенных параметров и позволяет перейти к заданию графиков другого типа.

При проведении двухфакторного дисперсионного анализа, как в нашем примере, достаточно построить один график. Для этого надо задать сначала имя фактора для горизонтальной оси (Horisontal Axis). Обычно это тот фактор, который имеет больше градаций, — в нашем случае это фактор хобби. Затем необходимо задать имя фактора для отдельных линий па графике (Separate Lines), — в пашем случае это фактор пол. В результате будет построен график с двумя ломаными линиями,

каждая из которых соединяет три точки. Вертикальная ось этого графика будет соответствовать величине средних значений, горизонтальная ось — переменной хобби, а две отдельные линии — переменной пол.

Для трехфакториого анализа подобный фафик позволит интерпретировать только двухфакторное взаимодействие. Для интерпретации трехфакторного взаимодействия необходимо градациям третьего фактора поставить в соответствие разные диаграммы (Separate Plots). Например, если бы третьим фактором в нашем примере был фактор класс, то задание этого фактора как имени для разных фафиков привело бы к построению трех фафиков — по одному для каждого класса, каждый из которых соответствовал бы двухфакторному взаимодействию переменных пол и хобби.

Пошаговый алгоритм, приведенный ниже, повторяет действия предшествующего шага, но добавляет в окно вывода график средних значений, позволяющий интерпретировать взаимодействие факторов пол и класс.

**Шаг 5б** После выполнения шага 4 должно быть открыто диалоговое окно Univariate (Одномерный анализ), показанное на рис. 14.1. Если вы уже успели поработать с этим окном, щелкните на кнопке Reset (Сброс).

1. Задайте переменные для анализа. Для этого щелкните на переменной отметка2, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в поле Dependent Variable (Зависимая переменная); щелкните на переменной пол, чтобы выделить ее, а затем — на второй сверху кнопке со стрелкой, чтобы переместить переменную в список Fixed Factor(s) (Постоянные факторы). Повторите то же действие для переменной хобби. Щелкните па переменной тест_ср, чтобы выделить ее, а затем — на второй снизу кнопке со стрелкой, чтобы переместить переменную в список Covariates (Ковариаты)
2. Задайте параметры результатов. Для этого щелкните па кнопке Options (Параметры), чтобы открыть диалоговое окно Univariate: Options (Одномерный анализ: Параметры), и установите флажки Descriptive statistics (Описательные статистики) и Estimates of effect size (Оценка величины эффекта), а затем щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Univariate (Одномерный анализ).
3. Щелкните на кнопке Graphs (Графики), чтобы открыть диалоговое окно Univariate: Profile Plots (Одномерный анализ: Диаграммы профилей).
4. Выделите в списке Factors (Факторы) переменную хобби и щелчком на верхней кнопке со стрелкой перенесите ее в поле Horisontal Axis (Горизонтальная ось).
5. Выделите в списке Factors (Факторы) переменную пол и щелчком на второй сверху кнопке со стрелкой перенесите ее в поле Separate Lines (Отдельные линии).
6. Подтвердите правильность введенных параметров щелчком на кнопке Add (Добавление) — в нижнем поле появится строка хобби*пол, обозначающая тип графика средних.
7. Щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Univariate (Одномерный анализ).
8. Щелкните на кнопке OK, чтобы открыть окно вывода.

После выполнения шага 5 программа автоматически активизирует окно вывода. Для просмотра результатов вы при необходимости можете воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

# Печать результатов и выход из программы

Ниже описана типичная процедура печати результатов статистического анализа (или нескольких анализов). После выполнения шага 5 должно быть открыто окно вывода.

**Шаг 6** В окне вывода укажите фрагменты, выводимые на печать (см. раздел «Окно вывода» в главе 2), в меню File (Файл) выберите команду Print (Печать), при необходимости задайте параметры печати и щелкните на кнопке OK.

Последнее, что необходимо сделать после завершения исследования и печати результатов, — это выйти из программы SPSS.

**Шаг 7** Для выхода из программы в меню File (Файл) выберите команду Exit (Выход).

Иногда после выполнения команды Exit (Выход) на экране могут появляться небольшие диалоговые окна с вопросом о необходимости сохранения сделанных в файлах изменений и кнопками, описывающими возможные варианты ответа. Для завершения работы просто щелкайте на соответствующих кнопках.

# Представление результатов

## Двухфакторный дисперсионный анализ

На рис. 14.4 приведены фрагменты выводимых данных, сгенерированные программой после выполнения шага 5.

В таблице Descriptive Statistics (Описательные статистики) содержатся характеристики всех выборок, образованных каждой из независимых переменных в отдельности (2 выборки по признаку пола и 3 выборки по переменной хобби), а также при пересечении градаций независимых переменных (3 × 2 = 6 выборок). Таблица Tests of Between-Subjects Effects (Критерии для межгрупповых эффектов) содержит результаты проверки трех основных гипотез двухфакторного дисперсионного анализа.

- Переменная пол оказывает статистически достоверное влияние на распределение зависимой переменной отметка2 (средние значения для юношей и девушек составили соответственно 4,13 и 4,28, $F = 4{,}74$, $p = 0{,}032$).

**Descriptive Statistics**

Dependent Variable: ОТМЕТКА2

| ПОЛ | ХОББИ | Mean | Std. Deviation | N |
|---|---|---|---|---|
| ЖЕН | спорт | 4.23 | .243 | 15 |
| | компьютер | 4.25 | .286 | 19 |
| | искусство | 4.33 | .271 | 27 |
| | Total | 4.28 | .269 | 61 |
| МУЖ | спорт | 4.00 | .259 | 18 |
| | компьютер | 4.27 | .224 | 18 |
| | искусство | 4.10 | .173 | 3 |
| | Total | 4.13 | .266 | 39 |
| Total | спорт | 4.10 | .273 | 33 |
| | компьютер | 4.26 | .254 | 37 |
| | искусство | 4.31 | .270 | 30 |
| | Total | 4.22 | .276 | 100 |

Tests of Between-Subjects **Effects**

Dependent Variable: ОТМЕТКА2

| Source | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|---|---|---|---|---|---|---|
| Corrected Model | 1.28[a] | 5 | .26 | 3.85 | .003 | .170 |
| Intercept | 1054.38 | 1 | 1054.38 | 15849.4 | .000 | .994 |
| ПОЛ | .32 | 1 | .32 | 4.74 | .032 | .048 |
| ХОББИ | .36 | 2 | .18 | 2.74 | .070 | .055 |
| ПОЛ * ХОББИ | .30 | 2 | .15 | 2.28 | .108 | .046 |
| Error | 6.26 | 94 | .07 | | | |
| Total | 1788.80 | 100 | | | | |
| Corrected Total | 7.54 | 99 | | | | |

a. R Squared = .170 (Adjusted R Squared = .126)

**Рис. 14.4.** Фрагменты окна вывода после выполнения шага 5

- Переменная хобби не оказывает статистически значимого влияния па распределение зависимой переменной отметка2 (средние значения для групп 1, 2 и 3 составили, соответственно, 4,10, 4,26 и 4,31, $F = 2{,}74$, $p = 0{,}070$).
- Не обнаружено статистически достоверного взаимодействия между независимыми переменными пол и хобби ($F = 2{,}28$, $p = 0{,}108$).

Взаимодействие независимых переменных более подробно обсуждается в последнем разделе этой главы, посвященном использованию графиков для интерпретации взаимодействий.

## Влияние ковариаты

На рис. 14.5 приведен фрагмент выводимых данных, сгенерированных программой после выполнения шага 5а.

Включение ковариаты не меняет описательные статистики, но вносит некоторые изменения в результаты проверки гипотез дисперсионного анализа, как видно из таблицы Tests of Between-Subjects Effects (Критерии для межгрупповых эффектов).

**Tests of Between-Subjects Effects**

Dependent Variable: ОТМЕТКА2

| Source | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|---|---|---|---|---|---|---|
| Corrected Model | 1.64[a] | 6 | .27 | 4.306 | .001 | .217 |
| Intercept | 23.77 | 1 | 23.77 | 374.9 | .000 | .801 |
| ТЕСТ_СР | .36 | 1 | .36 | 5.615 | .020 | .057 |
| ПОЛ | .23 | 1 | .23 | 3.610 | .061 | .037 |
| ХОББИ | .28 | 2 | .14 | 2.242 | .112 | .046 |
| ПОЛ * ХОББИ | .39 | 2 | .20 | 3.088 | .050 | .062 |
| Error | 5.90 | 93 | .06 | | | |
| Total | 1788.80 | 100 | | | | |
| Corrected Total | 7.54 | 99 | | | | |

a. R Squared = .217 (Adjusted R Squared = .167)

**Рис. 14.5.** Фрагмент окна вывода после выполнения шага 5а

Ковариата оказывает влияние па суммы квадратов и, как следствие, на число степеней свободы (в некоторых случаях), средние квадраты, величины *F* и $\eta^2$, а также значения *p*-уровней. Если это влияние существенно, то, как правило, значения среднего квадрата и *F*-критерия увеличиваются, а соответствующие уровни значимости уменьшаются.

Обратите внимание, что ковариата тест_ср оказывает значительное влияние па разброс зависимой переменной отметка2: значение $\eta^2$ составляет 0,057, то есть 5,7 % дисперсии переменной отметка2 обусловлено влиянием ковариаты. Дисперсия скорректированной модели представляет собой сумму всех сумм квадратов дисперсий, обусловленных влияниями независимых переменных и их взаимодействий.

Сравните между собой результаты дисперсионного анализа с ковариатой и без ковариаты. Благодаря тому что ковариата «съедает» часть дисперсии зависимой переменной, для взаимодействия факторов *F*-критерий увеличивается, а *p*-уровень уменьшается. В результате взаимодействие факторов становится статистически достоверным. Разумеется, подобный результат проявляется не всегда, и вы можете убедиться в этом на примере строк пол и хобби: в них введение ковариаты уменьшило *F*-критерии и увеличило *p*-уровни. В итоге снизилась статистическая значимость эффекта фактора пол.

Таким образом, двухфакторный дисперсионный анализ с зависимой переменной отметка2, независимыми переменными пол и хобби и ковариатой тест_ср дал следующие результаты:

- Ковариата тест_ср оказывает статистически достоверное влияние на зависимую переменную отметка2 ($F$ - 5,615, $p$ - 0,020).
- Переменная пол оказывает влияние на распределение зависимой переменной отметка2 лишь на уровне статистической тенденции ($F = 3,61$, $p = 0,061$).
- Переменная хобби не оказывает статистически значимого влияния па распределение зависимой переменной отметка2 ($F$ - 2,24, $p$ = 0,112).

- Обнаружено статистически достоверное взаимодействие между независимыми переменными пол и хобби ($F$ - 3,088, $p$ - 0,050).

Отметим, что не очень удобно интерпретировать взаимодействие факторов по средним значениям из таблицы Descriptive Statistics (Описательные статистики). Для этого лучше использовать графики средних значений, полученные после выполнения шага 5б.

# Терминология, используемая при выводе

Ниже дана трактовка терминов, используемых программой в окне вывода.

- Sum of squares (Сумма квадратов) — сумма квадратов отклонений от среднего значения. Различают несколько видов сумм квадратов. Скорректированная сумма квадратов модели (Corrected Model) учитывает отклонения, обусловленные независимыми переменными и их взаимодействием, а скорректированная полная сумма квадратов (Corrected Total) характеризует всю дисперсию распределения. Каждому фактору и их взаимодействию соответствует своя величина суммы квадратов. Остаточная сумма квадратов (Error) учитывает ту часть отклонений, которая не обусловлена влияниями со стороны независимых переменных или их взаимодействий.
- df (Число степеней свободы) — для ковариаты это число равно 1, для факторов — числу уровней фактора, уменьшенному на 1, для переменной хобби — 2 (то есть 3 – 1), для переменной пол — 1 (то есть 2 – 1). Для двух взаимодействующих факторов это число равно произведению степеней свободы факторов: для факторов хоббихпол оно равно 2 (то есть 2 x 1). Число степеней свободы скорректированной модели равно сумме степеней свободы независимых переменных и их взаимодействия (1 + 1 + 2 + 2 = 6). Число степеней свободы ошибки (остатка) равно разности числа объектов и числа степеней свободы скорректированной модели, уменьшенной на единицу (100 - 6 - 1 = 93). Скорректированное суммарное число степеней свободы равно числу объектов, уменьшенному на единицу (100 - 1 = 99).
- Mean square (Средний квадрат) — отношение суммы квадратов к числу степеней свободы.
- F (F-критерий) — отношение среднего квадрата независимой переменной или взаимодействия переменных к среднему квадрату остатка.
- Sig. (Значимость) — вероятность того, что различие (влияние) является случайным.
- Partial Eta Squared (Частичный квадрат эта) — оценка величины эффекта ($\eta^2$). Под эффектом понимается вклад независимой переменной или взаимодействия переменных в разброс значений зависимой переменной.

# Использование графиков для интерпретации взаимодействий

На рис. 14.6 приведен график средних значений, который генерируется программой после выполнения шага 5б.

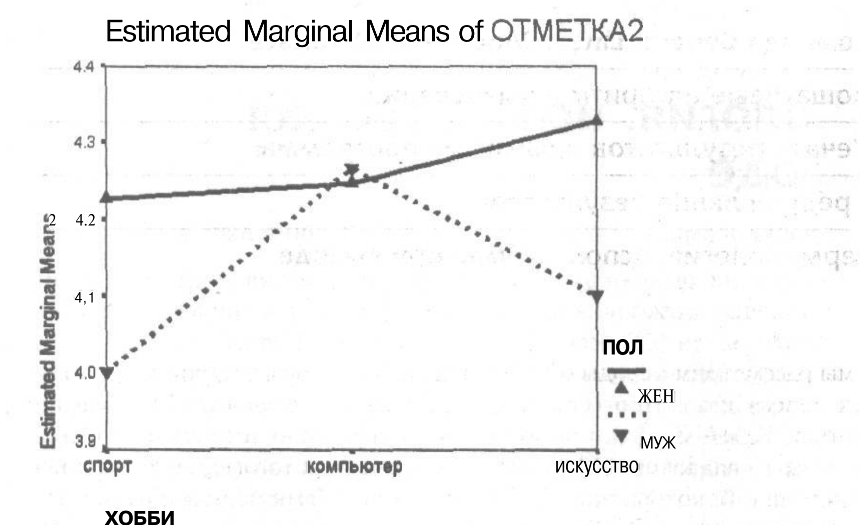



**Рис. 14.6.** График средних значений, генерируемый после выполнения шага 5б

График позволяет без труда интерпретировать взаимодействие факторов: влияние фактора хобби на переменную отметка2 проявляется по-разному на разных градациях переменной пол. Так, средний балл отметок для девушек примерно одинаков вне зависимости от внешкольных увлечений. Для юношей средний балл отметок значительно различается в зависимости от внешкольных увлечений: наиболее высокий средний балл имеют школьники, которые увлекаются компьютером.

Взаимодействие факторов на графике выглядит как заметное различие формы профилей соответствующих линий. Чем ближе по форме профили линий, тем меньше взаимодействие факторов. Но интерпретации подлежат только те графики, которые соответствуют статистически достоверному взаимодействию факторов.

Если бы проводился трехфакторный анализ, то интерпретация трехфакторного взаимодействия требовала бы построения нескольких графиков. Так, при добавлении фактора класс потребовалось бы построение трех графиков двухфакторного взаимодействия: по одному для каждой градации фактора класс. Попробуйте самостоятельно провести трехфакторный дисперсионный анализ, добавив при выполнении инструкций шага 5б переменную класс в качестве третьего фактора и при задании параметров фафиков средних значений в диалоговом окне Univariate: Profile Plots (Одномерный анализ: Диаграммы профилей) указав эту переменную в поле Separate Plots (Отдельные диаграммы).

За инструкциями по редактированию получаемых графиков обратитесь к главе 5.

# 15 Многомерный дисперсионный анализ



В этой главе мы рассмотрим методы обработки данных, которые содержат несколько зависимых переменных: многомерный дисперсионный анализ (Multivariate Analysis Of Variances, MANOVA) и многомерный ковариационный анализ (Multivariate Analysis Of Covariance, MANCOVA). Команды многомерного анализа, входящие в подменю General Linear Model (Общая линейная модель), относятся к наиболее сложным командам в SPSS[1]. Помимо команды Multivariate (Многомерный анализ), предназначенной для проведения анализов MANOVA и MANCOVA, к командам многомерного анализа относится и команда Repeated Measures (Повторные измерения), позволяющая провести анализ MANOVA с повторными измерениями. Первые два типа многомерного анализа будут рассмотрены в этой главе, третий — в главе 16.

Ввиду значительной сложности мы упомянем лишь наиболее понятные и широко используемые параметры команды Multivariate (Многомерный анализ). Фактически, критерии MANOVA и MAVCOVA являются ни чем иным, как расширением дисперсионного анализа (ANOVA), рассмотренного нами ранее (см. главы 13-14); если вы не знакомы с дисперсионным анализом, рекомендуем, прежде чем читать дальше, обратиться к указанным главам.

Как упоминалось в предыдущих главах, $t$-критерий для двух выборок позволяет выяснить, существуют ли различия между двумя средними значениями для этих выборок. Эту простейшую ситуацию (единственная независимая переменная с двумя градациями и одна зависимая переменная метрического типа) можно последовательно усложнить тремя способами:

- ввести в рассмотрение независимую переменную с более чем двумя градациями — в такой ситуации применяется однофакторный дисперсионный анализ;

[1] Для выполнения многомерного анализа требуется полная установка пакета SPSS, включая модули Advanced Models (Дополнительные модели) и Regression Models (Регрессионные модели).

- ввести не одну, а несколько независимых переменных — для этого предназначен многофакторный дисперсионный анализ;
- задействовать ковариаты.

Во всех трех случаях зависимая переменная остается единственной и имеет метрический тип. Тем не менее существуют задачи, в которых требуется учитывать не одну, а несколько зависимых переменных. В этой главе мы займемся рассмотрением проблемы проведения анализа с участием более чем одной зависимой переменной; при этом мы не станем усложнять требования к независимым переменным.

Обратимся к файлу данных ex01.sav. Представим себе, что нам необходимо сравнить мужчин и женщин (переменная пол) одновременно по всем пяти показателям теста (переменные тест1, ..., тест5). В подобной ситуации одним из возможных подходов является пятикратное применение $t$-критерия или однофакторного дисперсионного анализа (эти методы эквивалентны, поскольку $t^2 = F$). Очевидным достоинством такого решения является простота и ясность, однако нельзя не заметить и двух недостатков: во-первых, при неоднократном применении статистического критерия (в данном случае пятикратном) увеличивается вероятность ошибки, то есть вероятность случайности общего результата исследования; во-вторых, если между зависимыми переменными имеется некоторая корреляция (а в рассматриваемом случае она есть), то результат, полученный в отношении каждой из этих переменных в отдельности, не способен отразить этот важный факт.

Описанные недостатки привели к усовершенствованию как $t$-критерия, так и дисперсионного анализа: первый был расширен с помощью критерия Хотеллинга (Hotelling), а вместо второго стал использоваться многомерный дисперсионный анализ (MANOVA). Оба типа анализа реализуются командой Multivariate (Многомерный анализ). Кроме того, эта команда позволяет проводить многомерный ковариационный анализ (MANCOVA), учитывающий влияние ковариат. Особенностью всех типов многомерного анализа является то, что они обрабатывают все зависимые переменные одновременно. В примерах этой главы показано, каким образом исследовать структуру изменений зависимых переменных путем применения серий одномерных $F$-критериев или серий множественных сравнений постфактум.

Применяя MANOVA в отношении множества зависимых переменных, следует помнить, что линейная функциональная связь между ними недопустима. Иными словами, следует избегать применения анализа MANOVA к тем зависимым переменным, корреляция между которыми близка к 1.

Как и в случае одномерного дисперсионного анализа, в MANOVA для определения значимости различий между группами используются $F$-критерий. Отметим, что многомерный дисперсионный анализ, как и одномерный, позволяет оценить влияния не только отдельных независимых переменных (главных эффектов), но и их взаимодействий. Поскольку мы рассматриваем наличие нескольких зависи-

мых переменных, $F$-статистика носит многомерный характер, и для ее формирования используется матричная алгебра.

Для иллюстрации многомерного дисперсионного анализа используется файл данных ex02.sav. Этот файл содержит гипотетические результаты эксперимента по изучению эффективности запоминания слов в зависимости от частоты их встречаемости и от интонации, с которой они предъявлялись (зачитывались). Ряды из 24 не связанных по смыслу слов одинаковой длины зачитывались 20 испытуемым. Сразу после предъявления испытуемых просили воспроизвести эти слова. Подсчитывалось количество правильно воспроизведенных слов из начала ряда — первых 8 слов (переменная начало1), из середины ряда (переменная средн1) и из конца ряда — завершающих 8 слов (переменная конец1). Переменная инт соответствует делению испытуемых на две группы: первой (инт - 1) все слова читались с одинаковой интонацией; второй (инт - 2) середина ряда интонационно выделялась. Переменной част соответствует деление испытуемых па тех, кому предъявлялся ряд часто встречающихся слов (част - 1), и тех, кому предъявлялся ряд редко встречающихся слов (част - 2). Кроме того, каждый предъявляемый ряд составлялся из слов одинаковой эмоциональной значимости — ряды различались и по этой переменной (знач — количественная переменная, отражающая эмоциональную значимость). Таким образом, данные позволяют проверить гипотезы о влиянии двух независимых переменных (инт — интонация, част — частота) на три зависимые переменные (начало1, средн1, конец1). Переменная знач используется как ковариата, иными словами, ее влияние при проверке указанных гипотез будет исключено.

Для многомерного анализа необходимо иметь как минимум две зависимые переменные (иначе анализ не является многомерным) и как минимум одну независимую переменную. Теоретически количество зависимых и независимых неременных не ограничено, однако на практике объем выборки диктует необходимость существенного ограничения их числа. Использование ковариат, разумеется, не является обязательным.

# Пошаговые алгоритмы вычислений

При проведении многомерного дисперсионного анализа сначала выполняются три подготовительных шага.

**Шаг 1** Создайте новый файл данных или подготовьте существующий. Все необходимые действия описаны в главе 3.

**Шаг 2** Запустите программу SPSS при помощи значка на рабочем столе или команды Пуск ▸ Программы ▸ SPSS for Windows ▸ SPSS 11.5 for Windows (Start ▸ Programs ▸ SPSS for Windows ▸ SPSS 11.5 for Windows) главного меню Windows. В открывшемся после запуска программы диалоговом окне SPSS for Windows щелкните на кнопке Cancel (Отмена).

После выполнения этого шага на экране появится окно редактора данных SPSS.

**Шаг 3** Откройте файл данных, с которым вы намерены работать (в пашем случае это файл ex02.sav). Если он расположен в текущей папке, то выполните следующие действия:

1. Выберите в меню File (Файл) команду Open ▸ Data (Открытие ▸ Данные) или щелкните на кнопке Open File (Открытие файла) панели инструментов.
2. В открывшемся диалоговом окне дважды щелкните на имени ex02.sav или введите его с клавиатуры и щелкните на кнопке OK.

Независимо от того, открыта программа SPSS только что или какие-то процедуры уже выполнялись, в верхней части главного окна должна присутствовать строка меню (она показана ниже). Пока строка меню присутствует па экране, доступны все команды анализа данных. При этом окно с данными видеть не обязательно.

File Edit View Data Transform Analyze Graphs Utilities Window Help

При работе с таблицами результатов или при редактировании диаграмм некоторые пункты строки меню могут исчезать или меняться. Чтобы вернуться *к* главному окну и стандартной строке меню, щелкните на кнопке свертывания или восстановления текущего окна.

После завершения шага 3 на экране должно присутствовать окно редактора данных со строкой меню.

**Шаг 4** В меню Analyze (Анализ) выберите команду General Linear Model ▸ Multivariate (Общая линейная модель ▸ Многомерный анализ). Откроется диалоговое окно Multivariate (Многомерный анализ), показанное на рис. 15.1.

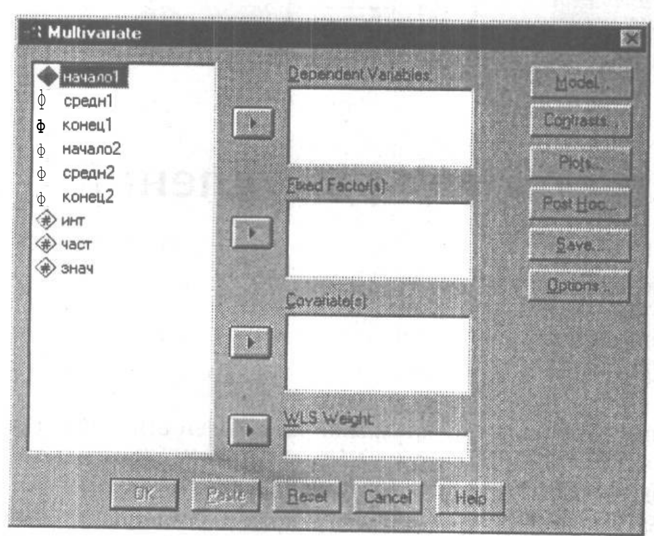



Рис. 15.1. Диалоговое окно Multivariate

Окно Multivariate (Многомерный анализ) позволяет задавать переменные для многомерного анализа, а также управлять диалоговыми окнами, предназначенными для определения параметров анализа. Те переменные, которые будут использоваться в качестве зависимых, необходимо поместить в список Dependent Variables (Зависимые переменные), а независимые переменные — в список Fixed Factor(s) (Постоянные факторы).

Как и прежде, для перемещения переменных из исходного списка в целевой используется одна из кнопок со стрелками. Мы будем использовать в наших примерах анализа 3 зависимые и 2 независимые переменные из файла ex02.sav. При желании в анализ могут быть введены одна или несколько ковариат с помощью списка Covariate(s) (Ковариаты).

В правой части диалогового окна расположены 6 кнопок, предназначенные для настройки параметров анализа. Использование кнопок Contrasts (Контрасты) и Save (Сохранить) не рассматривается, поскольку настройка соответствующих параметров требует очень высокой квалификации исследователя.

В большинстве случаев требуется многомерный анализ как для каждой из независимых переменных, так и для всех вариантов их взаимодействий (полнофакторная модель). Полнофакторная модель является моделью анализа, используемой по умолчанию. С другой стороны, иногда полезно исключить из рассмотрения часть взаимодействий или отдельные переменные. Для управления моделью анализа используется кнопка Model (Модель) в диалоговом окне Multivariate (Многомерный анализ). При щелчке на ней открывается диалоговое окно Multivariate: Model (Многомерный анализ: Модель), представленное на рис. 15.2.

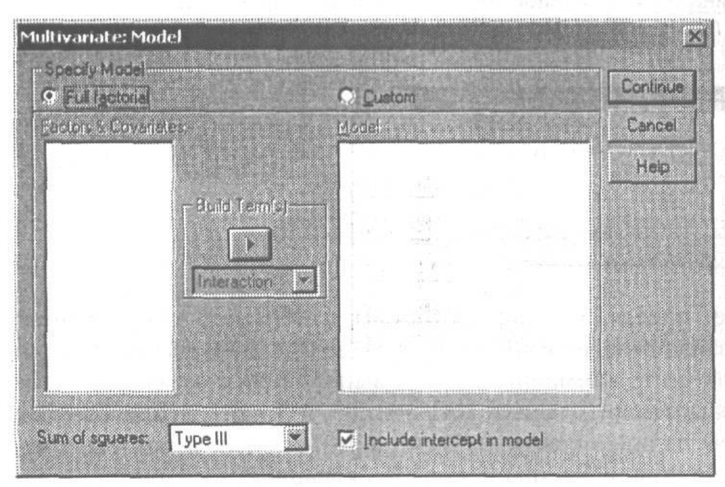



**Рис. 15.2.** Диалоговое окно Multivariate: Model

По умолчанию в группе Specify Model (Выбор модели) установлен переключатель Full factorial (Полнофакторная модель), однако вы можете установить переключатель Custom (Настройка), а затем с помощью кнопки с направленной

вправо стрелкой переместить в список Model (Модель) только те переменные из списка Factors & Covariates (Факторы и ковариаты), которые вы хотите использовать.

Диалоговое окно Multivariate: Model (Многомерный анализ: Модель) с помощью списка Sum of squares (Сумма квадратов) позволяет также задать вариант расчета суммы квадратов. По умолчанию выбран вариант Type I I I (Тип III), подходящий для большинства ситуаций. Если исходные данные содержат пропущенные значения, иногда выбирают пункт Type IV (Тип IV).

При интерпретации результатов многомерного анализа зачастую удобно иметь перед глазами графическое представление средних значений зависимых переменных, определяемых различными комбинациями градаций независимых переменных (факторов). Для построения подобных графиков используется кнопка Plots (Диаграммы), назначение которой идентично назначению одноименной кнопки в окне Univariate (Одномерный анализ), используемом при проведении одномерного дисперсионного анализа (см. главу 14). При щелчке на ней открывается диалоговое окно Multivariate: Profile Plots (Многомерный анализ: Диаграммы профилей), представленное на рис. 15.3.

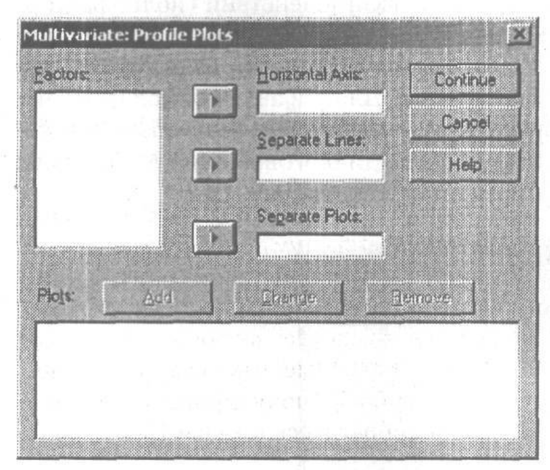



**Рис. 15.3.** Диалоговое окно Multivariate: Profile Plots

Данное окно позволяет определить, каким образом отображать тот или иной фактор. Для каждой зависимой переменной в выводимые данные включаются отдельные графики. Обратите внимание, что при выполнении анализа с участием ковариат вместо фактических данных отображаются приближения средних значений, учитывающие наличие ковариат. В подобных случаях может быть полезным получение изображений без использования ковариат.

В левой части окна Multivariate: Profile Plots (Многомерный анализ: Диаграммы профилей) находится список Factors (Факторы), содержащий те факторы, которые присутствуют в модели анализа. Окно содержит три поля: Horizontal Axis (Горизонтальная ось), Separate Lines (Отдельные линии) и Separate Plots (Отдельные диаграммы).

В этих полях можно указать соответственно фактор, категории которого должны отображаться по горизонтальной оси, фактор, который должен отображаться в виде отдельных линий (по одной линии для каждой категории), и фактор, для каждого уровня которого должна строиться отдельная диаграмма.

Из перечисленных полей обязательным для заполнения является только иоле Horizontal Axis (Горизонтальная ось), а чтобы ввести переменную в поле Separate Plots (Отдельные диаграммы), необходимо сначала заполнить поле Separate Lines (Отдельные линии). После заполнения этих полей щелкните на кнопке Add (Добавление). Соответствующая информация появится в списке в нижней части диалогового окна. Когда все параметры графика будут заданы, щелкните на кнопке Continue (Продолжить), чтобы вернуться в основное диалоговое окно.

Графическая интерпретация весьма полезна для оценки влияния независимых переменных и их взаимодействий. Для более детального анализа требуются множественные (парные) сравнения постфактум. Множественные сравнения проводятся в случае установления статистически достоверного влияния независимой переменной и позволяют определять, между какими именно градациями независимой переменной имеются различия. Подобные сравнения проводятся для каждой из зависимых переменных и каждого выбранного фактора. Множественные сравнения неприменимы при использовании ковариат.

Для проведения парных сравнений применяется кнопка Post Hoc (Постфактум). При щелчке на ней открывается диалоговое окно Multivariate: Post Hoc Multiple Comparisons for Observed Means (Многомерный анализ: Множественные сравнения средних значений постфактум), представленное на рис. 15.4. Это окно идентично диалоговому окну One-Way ANOVA: Post Hoc Multiple Comparisons (Однофакторный дисперсионный анализ: Множественные сравнения постфактум), которое подробно описано в главе 13. Отличается окно для многомерного анализа только наличием списков Factor(s) (Факторы) и Post Hoc Tests for (Критерии постфактум для). В список Post Hoc Tests for (Критерии постфактум для) следует включить факторы, для которых будут проводиться сравнения. Все доступные факторы перечислены в списке Factor(s) (Факторы). За описанием наиболее популярных критериев парных сравнений и результатов их применения обратитесь к главе 13.

При щелчке на кнопке Options (Параметры) открывается диалоговое окно Multivariate: Options (Многомерный анализ: Параметры), представленное на рис. 15.5. По умолчанию все флажки в этом окне сброшены.

Если выделить нужные пункты в списке Factor(s) and Factor Interactions (Факторы и взаимодействия факторов) и переместить их в список Display Means For (Отображать средние для), то в выводимые данные будут включены средние значения каждой из зависимых переменных для всех уровней соответствующих факторов. Если анализ проводится без ковариат, включаемые средние значения будут фактическими средними значениями данных; в противном случае средние значения будут подсчитаны с учетом влияния ковариат. Пункт (Overall) (Все) в списке Factor(s) and Factor Interactions (Факторы и взаимодействия факторов) позволяет вычислить средние значения для всех ячеек. Если установить флажок Compare

main effects (Сравнивать воздействия факторов), будет проведена серия парных сравнений постфактум средних значений ячеек для каждого из факторов. Вы можете использовать либо критерий Бонферрони (Bonferroni), либо более «консервативный» критерий Сидака (Sidak). Другие критерии задаются при помощи описанного ранее диалогового окна Multivariate: Post Hoc Multiple Comparisons for Observed Means (Многомерный анализ: Множественные сравнения постфактум наблюдаемых средних значений) при условии, что анализ проводится без ковариат.

**Рис. 15.4.** Диалоговое окно Multivariate: Post Hoc Multiple Comparisons for Observed Means

**Рис. 15.5.** Диалоговое окно Multivariate: Options

Мы опустим описания большей части флажков и упомянем лишь те, которые используются наиболее часто.

- Descriptive Statistics (Описательные статистики) — вычисляются средние значения и стандартные отклонения каждой зависимой переменной для каждой ячейки.
- Estimates of effect size (Оценки величины эффекта) — вычисляется значение $\eta^2$, характеризующее величину воздействия фактора на зависимую переменную и показывающее, какая доля общей дисперсии обусловлена данным фактором.
- Parameter estimates (Оценки параметров) — вычисляются приближения и соответствующие значимости для всех факторов и ковариат модели. Это особенно удобно в случаях, когда в модель включена хотя бы одна ковариата.
- Homogeneity tests (Критерии однородности) — дисперсии для всех ячеек проверяются на равенство (однородность).

Поле Significance level (Уровень значимости) позволяет задать величину а для анализа. По умолчанию задан критический уровень значимости 0,05; если вы измените это значение, SPSS пересчитает соответствующие доверительные интервалы.

Как было сказано выше, остальные флажки диалогового окна Multivariate: Options (Многомерный анализ: Параметры) в этой книге не рассматриваются. При необходимости вы можете найти их описания в руководстве пользователя SPSS. Кроме того, некоторая дополнительная информация содержится в последней главе.

В рассматриваемом далее примере проводится многомерный дисперсионный анализ (MANOVA) с тремя зависимыми переменными начало1, средн1, конец1 и двумя независимыми переменными (факторами) инт и част.

**Шаг 5** После выполнения предыдущего шага у вас должно быть открыто диалоговое окно Multivariate (Многомерный анализ), показанное на рис. 15.1. Если вы уже успели поработать с этим окном, щелкните на кнопке Reset (Сброс).

1. Щелкните на переменной начало1, нажмите клавишу Shift и, не отпуская ее, щелкните на переменной конец1. В результате окажутся выделенными переменные начало1, средн1, конец1.
2. Щелкните на верхней кнопке со стрелкой, чтобы переместить выделенные переменные в список Dependent Variables (Зависимые переменные).
3. Щелкните сначала на переменной инт, чтобы выделить ее, а затем — на второй сверху кнопке со стрелкой, чтобы переместить переменную в список Fixed Factor(s) (Постоянные факторы).
4. Повторите предыдущее действие для переменной част.
5. Щелкните на кнопке OK, чтобы открыть окно вывода.

В следующем примере мы включим в анализ ковариату знач.

**Шаг 5а** После выполнения шага 4 у вас должно быть открыто диалоговое окно окно Multivariate (Многомерный анализ), показанное на рис. 15.1. Если вы уже успели поработать с этим окном, щелкните на кнопке Reset (Сброс).

1. Щелкните на переменной начало1, нажмите клавишу Shift и, не отпуская ее, щелкните на переменной конец1. В результате окажутся выделенными переменные начало1, средн1, конец1.
2. Щелкните на верхней кнопке со стрелкой, чтобы переместить выделенные переменные в список Dependent Variables (Зависимые переменные).
3. Щелкните сначала на переменной инт, чтобы выделить ее, а затем — на второй сверху кнопке со стрелкой, чтобы переместить переменную в список Fixed Factor(s) (Постоянные факторы).
4. Повторите предыдущее действие для переменной част.
5. Щелкните сначала на переменной знач, чтобы выделить ее, а затем — на третьей сверху кнопке со стрелкой, чтобы переместить переменную в список Covariate(s) (Ковариаты).
6. Щелкните на кнопке OK, чтобы открыть окно вывода.

В следующем примере демонстрируется использование нескольких описанных ранее параметров, генерирующих полезную статистическую и графическую информацию. Обратите внимание, что мы не включаем в данный анализ ковариату. Напоминаем, что для анализа с ковариатами парные сравнения неприменимы, поскольку все графики будут соответствовать не фактическим исходным данным, а данным, скорректированным влияниями ковариат. Кроме того, отметим, что парные сравнения имеют смысл лишь для факторов, имеющих более двух уровней; по этой причине в данном примере парные сравнения не применяются. Если же у вас какой-либо фактор имеет более двух уровней, воспользуйтесь кнопкой Post Hoc (Постфактум). Она действует практически идентично своему аналогу из диалогового окна One-Way ANOVA (Однофакторный дисперсионный анализ), поэтому при необходимости вернитесь к главе 13. Отличие лишь в том, что в данном случае следует указать каждый фактор, для которого необходимы парные сравнения.

**Шаг 5б** После выполнения шага 4 у вас должно быть открыто диалоговое окно Multivariate (Многомерный анализ), показанное па рис. 15.1. Если вы уже успели поработать с этим окном, щелкните на кнопке Reset (Сброс).

1. Щелкните на переменной начало1, нажмите клавишу Shift и, не отпуская ее, щелкните на переменной конец1. В результате окажутся выделенными переменные начало1, средн1, конец1.
2. Щелкните на верхней кнопке со стрелкой, чтобы переместить выделенные переменные в список Dependent Variables (Зависимые переменные).
3. Щелкните сначала на переменной инт, чтобы выделить ее, а затем — на второй сверху кнопке со стрелкой, чтобы переместить переменную в список Fixed Factor(s) (Постоянные факторы).

4. Повторите предыдущее действие для переменной част.
5. Щелкните на кнопке Plots (Диаграммы), чтобы открыть диалоговое окно Multivariate: Profile Plots (Многомерный анализ: Диаграммы профилей), показанное на рис. 15.3.
6. Щелкните сначала на переменной инт, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в поле Horizontal Axis (Горизонтальная ось).
7. Щелкните сначала на переменной част, чтобы выделить ее, а затем — на средней кнопке со стрелкой, чтобы переместить переменную в поле Separate Lines (Отдельные линии).
8. Щелкните сначала на кнопке Add (Добавление), чтобы зафиксировать заданные параметры построения диаграмм, а затем — на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Multivariate (Многомерный анализ).
9. Щелкните на кнопке Options (Параметры), чтобы открыть диалоговое окно Multivariate: Options (Многомерный анализ: Параметры), показанное на рис. 15.5.
10. Щелкните сначала на переменной инт, чтобы выделить ее, а затем — на кнопке со стрелкой, чтобы переместить переменную в список Display Means for (Отображать средние для).
11. Повторите предыдущее действие для переменной част и установите флажки Descriptive Statistics (Описательные статистики), Estimates of effect size (Оценивать величину воздействия), Homogeneity tests (Критерии однородности). Щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Multivariate (Многомерный анализ).
12. Щелкните на кнопке OK, чтобы открыть окно вывода.

После выполнения шага 5 программа автоматически активизирует окно вывода. Для просмотра результатов вы при необходимости можете воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

# Печать результатов и выход из программы

Ниже описана типичная процедура печати результатов статистического анализа (или нескольких анализов). После выполнения шага 5 должно быть открыто окно вывода.

**Шаг 6** В окне вывода укажите фрагменты, выводимые на печать (см. раздел «Окно вывода» в главе 2), в меню File (Файл) выберите команду Print (Печать), при необходимости задайте параметры печати и щелкните на кнопке OK.

Последнее, что необходимо сделать после завершения исследования и печати результатов, — это выйти из программы SPSS.

Шаг 7 Для выхода из программы в меню File (Файл) выберите команду Exit (Выход).

Иногда после выполнения команды Exit (Выход) на экране могут появляться небольшие диалоговые окна с вопросом о необходимости сохранения сделанных в файлах изменений и кнопками, описывающими возможные варианты ответа. Для завершения работы просто щелкайте на соответствующих кнопках.

# Представление результатов

На представленных далее рисунках приведены наиболее важные фрагменты выводимых данных, генерируемых программой при выполнении шагов 5 и 5б. Мы не приводим эти данные целиком, поскольку их объем составляет 13 страниц.

## Межгрупповые факторы

В таблице, представленной на рис. 15.6, перечислены все независимые переменные со своими уровнями и объемами выборок (N).

Between-Subjects **Factors**

| | | Value Label | N |
|---|---|---|---|
| ИНТ | 1.00 | нет | 10 |
| | 2.00 | есть | 10 |
| ЧАСТ | 1.00 | вч | 10 |
| | 2.00 | нч | 10 |

Рис. 15.6. Фрагмент окна вывода после выполнения шага 5 (межгрупповые факторы)

## Описательные статистики

Таблица описательных статистик, полученная в результате выполнения шага 5 (на рис. 15.7 представлен только ее фрагмент), содержит средние значения, стандартные отклонения и объемы выборок для всех элементов модели.

## Проверка ковариационных матриц на равенство

На рис. 15.8 приведены результаты теста различий ковариационных матриц. Эти результаты говорят о том, что ковариационные матрицы равны между собой (значение $p$-уровня значительно превышает 0,05). Если бы результаты теста были другими, то мы не могли бы ручаться за достоверность многомерного анализа, одним из условий проведения которого, как вы помните, является равенство ковариационных матриц.

**Descriptive Statistics**

| | ИНТ | ЧАСТ | Mean | Std. Deviation | N |
|---|---|---|---|---|---|
| НАЧАЛО1 | нет | вч | 5.0000 | 1.58114 | 5 |
| | | нч | 3.4000 | .89443 | 5 |
| | | Total | 4.2000 | 1.47573 | 10 |
| | есть | вч | 4.0000 | 1.58114 | 5 |
| | | нч | 2.6000 | 1.14018 | 5 |
| | | Total | 3.3000 | 1.49443 | 10 |
| | Total | вч | 4.5000 | 1.58114 | 10 |
| | | нч | 3.0000 | 1.05409 | 10 |
| | | Total | 3.7500 | 1.51744 | 20 |

**Рис. 15.7.** Фрагмент окна вывода после выполнения шага 5 (описательные статистики)

**Box's Test of** Equality of **Covariance** Matrices[a]

| | |
|---|---|
| Box's M | 9.520 |
| F | .339 |
| df1 | 18 |
| df2 | 904.638 |
| Sig. | .996 |

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.

a. Design: Intercept+ИНТ+ЧАСТ+ИНТ * ЧАСТ

**Рис. 15.8.** Фрагмент окна вывода после выполнения шага 5 (проверка ковариационных матриц на равенство)

## Многомерные критерии

Таблица, показанная на рис. 15.9, — результат выполнения шага 5 (без ковариаты). В ней содержится информация, которая является результатом проверки влияния факторов и их взаимодействий на зависимые переменные. Секция Intercept характеризует остаточную дисперсию (как правило, дисперсию ошибки). Как можно видеть, обнаружены статистически значимые главные эффекты факторов инт и част ($p < 0{,}01$), но не выявлено взаимодействие этих факторов. Если в анализе участвуют ковариаты, они также отображаются в этой таблице. Выполнив шаг 5а, вы можете убедиться, что ковариата знач, применяемая в нашем примере, оказывает значимое влияние на зависимые переменные и меняет статистическую значимость влияния факторов и их взаимодействия.

Далее дана трактовка терминов, используемых программой в окне вывода.

- Value (Значение) — в этом столбце представлены значения различных критериев для проверки влияния независимых переменных. Из всех критериев наиболее надежным считается критерий Пилая (Pillai).
- F (*F*-критерий) — оценка *F*-критерия.

**Multivariate Tests**

| Effect | | Value | F | Hypothesis df | Error df | Sig. |
|---|---|---|---|---|---|---|
| Intercept | Pillai's Trace | .976 | 192.6 | 3.000 | 14.0 | .000 |
| | Wilks' Lambda | .024 | 192.6 | 3.000 | 14.0 | .000 |
| | Hotelling's Trace | 41.277 | 192.6 | 3.000 | 14.0 | .000 |
| | Roy's Largest Root | 41.277 | 192.6 | 3.000 | 14.0 | .000 |
| ИНТ | Pillai's Trace | .896 | 40.01 | 3.000 | 14.0 | .000 |
| | Wilks' Lambda | .104 | 40.01 | 3.000 | 14.0 | .000 |
| | Hotelling's Trace | 8.573 | 40.01 | 3.000 | 14.0 | .000 |
| | Roy's Largest Root | 8.573 | 40.01 | 3.000 | 14.0 | .000 |
| ЧАСТ | Pillai's Trace | .494 | 4.552 | 3.000 | 14.0 | .020 |
| | Wilks' Lambda | .506 | 4.552 | 3.000 | 14.0 | .020 |
| | Hotelling's Trace | .975 | 4.552 | 3.000 | 14.0 | .020 |
| | Roy's Largest Root | .975 | 4.552 | 3.000 | 14.0 | .020 |
| ИНТ ' ЧАСТ | Pillai's Trace | .158 | .875 | 3.000 | 14.0 | .478 |
| | Wilks' Lambda | .842 | .875 | 3.000 | 14.0 | .478 |
| | Hotelling's Trace | .187 | .875 | 3.000 | 14.0 | .478 |
| | Roy's Largest Root | .187 | .875 | 3.000 | 14.0 | .478 |

**Рис. 15.9.** Фрагмент окна вывода после выполнения шага 5 (многомерные критерии)

- Hypothesis df (Гипотетическое число степеней свободы) — произведение числа зависимых переменных и уровней каждой независимой переменной, уменьшенных на 1. Для данного примера:

$$3 \times (2 - 1) \times (2 - 1) = 2 \times 1 \times 1 = 3.$$

- Error df (Ошибка числа степеней свободы) — величина, вычисляемая различными способами в зависимости от теста.
- Sig. (Значимость) — уровень значимости для соответствующего $F$-критерия.

## Критерий однородности дисперсий

Критерий однородности дисперсий (критерий Левина) позволяет проверить допущение о том, что дисперсии всех зависимых переменных одинаковы (рис. 15.10). Большие значения $p$-уровня для каждой из зависимых переменных не дают поводов для излишнего беспокойства относительно корректности дисперсионного анализа.

## Одномерные критерии

В дополнение к многомерным критериям для всех зависимых переменных, к каждой из них в отдельности применяется одномерный $F$-критерий. Несмотря на то что многомерный критерий, как было сказано ранее, является более совершенным, чем серия одномерных критериев, последние необходимы для интерпретации статистически достоверных результатов применения многомерного критерия. Таким образом, если многомерный критерий дает статистически значимый результат, для детальной интерпретации используют результаты применения одномерных критериев.

Levene's **Test of Equality of Error Variances**[a]

| | F | df1 | df2 | Sig. |
|---|---|---|---|---|
| НАЧАЛО1 | .611 | 3 | 16 | .617 |
| СРЕДН1 | .837 | 3 | 16 | .493 |
| КОНЕЦ1 | .782 | 3 | 16 | .521 |

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept+ИНТ+ЧАСТ+ИНТ * ЧАСТ

**Рис. 15.10.** Фрагмент окна вывода после выполнения шага 5 (критерий однородности дисперсий)

В нашем случае статистически достоверными являются результаты применения многомерных критериев для главных эффектов факторов инт и част. На рис. 15.11 приведен фрагмент таблицы с результатами одномерного анализа, относящимися к этим факторам. Как видно по этому фрагменту, влияние фактора инт статистически достоверно в отношении двух из трех зависимых переменных: средн1 и конец1. Влияние фактора част статистически достоверно в отношении всех зависимых переменных. Для содержательной интерпретации применения одномерных критериев в нашем случае можно обратиться к описательным статистикам (средним), а лучше — к графикам средних значений. Если бы факторы имели более двух градаций, то перед интерпретацией необходимо было бы применить метод парных сравнений постфактум. В случае если анализ включает ковариаты, то для каждой зависимой переменной одномерный критерий также будет учитывать влияние этих ковариат.

Ниже дана трактовка терминов, используемых программой в окне вывода.

- Sum of Squares (Сумма квадратов) — для факторов и их взаимодействий это межгрупповая сумма квадратов, то есть сумма квадратов разностей между главным средним значением и средними значениями каждой группы с весовыми коэффициентами, кратными объему соответствующих выборок.
- df (Число степеней свободы) — для факторов и их взаимодействий это произведение чисел уровней независимых переменных, уменьшенных на единицу; для ошибки — разность между числом всех объектов и числом степеней свободы для факторов и их взаимодействий, уменьшенная на единицу.
- Mean Square (Средний квадрат) — отношение суммы квадратов для данного фактора или взаимодействия к числу степеней свободы.
- F ($F$-критерий) — отношение гипотетического среднего квадрата к среднему квадрату ошибки.
- Sig. (Значимость) — $p$-уровень для соответствующего $F$-критерия.
- Partial Eta Squared (Частичный квадрат Эта) — значение ($\eta^2$), характеризующее меру воздействия на зависимую переменную и равное доле дисперсии, обусловленной соответствующим воздействием.

**Tests of** Between-Subjects **Effects**

| Source | Dependent Variable | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|---|---|---|---|---|---|---|---|
| Corrected Model | НАЧАЛО1 | 15.350 | 3 | 5.117 | 2.883 | .068 | .351 |
| | СРЕДН1 | 23.350 | 3 | 7.783 | 9.434 | .001 | .639 |
| | КОНЕЦ1 | 26.200 | 3 | 8.733 | 11.64 | .000 | .686 |
| Intercept | НАЧАЛО1 | 281.250 | 1 | 281.25 | 158.5 | .000 | .908 |
| | СРЕДН1 | 378.450 | 1 | 378.45 | 458.7 | .000 | .966 |
| | КОНЕЦ1 | 369.800 | 1 | 369.80 | 493.1 | .000 | .969 |
| ИНТ | НАЧАЛО1 | 4.050 | 1 | 4.050 | 2.282 | .150 | .125 |
| | СРЕДН1 | 14.450 | 1 | 14.450 | 17.52 | .001 | .523 |
| | КОНЕЦ1 | 16.200 | 1 | 16.200 | 21.60 | .000 | .574 |
| ЧАСТ | НАЧАЛО1 | 11.250 | 1 | 11.250 | 6.338 | .023 | .284 |
| | СРЕДН1 | 8.450 | 1 | 8.450 | 10.24 | .006 | .390 |
| | КОНЕЦ1 | 9.800 | 1 | 9.800 | 13.07 | .002 | .450 |
| ИНТ * ЧАСТ | НАЧАЛО1 | .050 | 1 | .050 | .028 | .869 | .002 |
| | СРЕДН1 | .450 | 1 | .450 | .545 | .471 | .033 |
| | КОНЕЦ1 | .200 | 1 | .200 | .267 | .613 | .016 |

**Рис. 15.11.** Фрагмент окна вывода после выполнения шага 5 (одномерные критерии)

## Оценки средних значений для уровней факторов

Таблицы, представленные на рис. 15.12, получены после выполнения шага 5б. Они содержат средние значения каждой зависимой переменной для каждой градации факторов. По этим данным можно дать подробную содержательную интерпретацию статистически достоверных результатов. Так, статистически достоверное влияние фактора инт на совокупность зависимых переменных было детализировано результатами применения одномерных критериев: фактор инт значимо влияет на переменные средн1 и конец1. Судя по средним значениям для этих переменных, интонационное выделение середины ряда слов приводит к тому, что эта часть ряда запоминается лучше, а конец ряда — хуже, чем без интонационного выделения. Влияние фактора част проявляется в том, что высокочастотные слова запоминаются лучше, чем низкочастотные, независимо от положения в ряду. Отсутствие статистически достоверного взаимодействия факторов инт и част свидетельствует о том, что влияние интонации на запоминание слов не зависит от частоты встречаемости запоминаемых слов (и наоборот, влияние частоты встречаемости слов на их запоминание не зависит от интонации, с которой они предъявлялись).

Взаимодействие факторов легче всего интерпретировать по графикам средних, которые получены после выполнения шага 5б. На рис. 15.13 приведен один из этих графиков для переменной средн1. Как показывает график, взаимодействие факторов весьма незначительно.

**Estimated Marginal Means**

**1. ИНТ**

| Dependent Variable | ИНТ | Mean | Std. Error | 95% Confidence Interval | |
|---|---|---|---|---|---|
| | | | | Lower Bound | Upper Bound |
| НАЧАЛО1 | нет | 4.200 | .421 | 3.307 | 5.093 |
| | есть | 3.300 | .421 | 2.407 | 4.193 |
| СРЕДН1 | нет | 3.500 | .287 | 2.891 | 4.109 |
| | есть | 5.200 | .287 | 4.591 | 5.809 |
| КОНЕЦ1 | нет | 5.200 | .274 | 4.619 | 5.781 |
| | есть | 3.400 | .274 | 2.819 | 3.981 |

2. ЧАСТ

| Dependent Variable | ЧАСТ | Mean | Std. Error | 95% Confidence Interval | |
|---|---|---|---|---|---|
| | | | | Lower Bound | Upper Bound |
| НАЧАЛО1 | вч | 4.500 | .421 | 3.607 | 5.393 |
| | нч | 3.000 | .421 | 2.107 | 3.893 |
| СРЕДН1 | вч | 5.000 | .287 | 4.391 | 5.609 |
| | нч | 3.700 | .287 | 3.091 | 4.309 |
| КОНЕЦ1 | вч | 5.000 | .274 | 4.419 | 5.581 |
| | нч | 3.600 | .274 | 3.019 | 4.181 |

**Рис. 15.12.** Фрагмент окна вывода после выполнения шага 5б (оценки средних значений для уровней факторов)

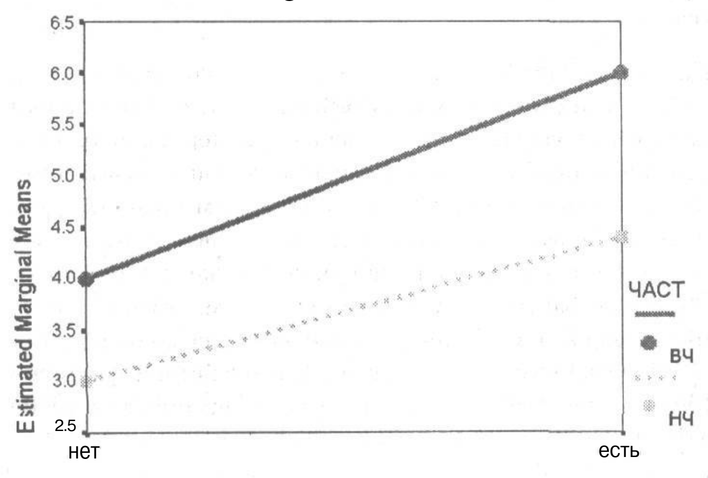



**Рис. 15.13.** Фрагмент окна вывода после выполнения шага 5б (график средних)

# 16 Дисперсионный анализ с повторными измерениями



В предыдущей главе рассматривалась ситуация, когда необходимо было анализировать влияние на совокупность зависимых переменных нескольких независимых переменных, каждая из которых могла иметь два или более уровней. В подобных случаях проводились межгрупповые сравнения: группам соответствовали уровни факторов и их сочетания. Такие факторы, разным уровням которых соответствуют разные группы объектов, называются межгрупповыми. Однако на практике нередко используются и внутригрупповые факторы, разным уровням которых соответствует одна и та же группа объектов. В этом случае говорят о внутригрупповой модели, с помощью которой исследуется одна и та же группа объектов для нескольких уровней независимой переменной.

В примерах этой главы мы будем использовать файл ex02.sav, содержащий гипотетические результаты эксперимента по изучению эффективности запоминания слов в зависимости от частоты их встречаемости и от интонации, с которой они предъявлялись (зачитывались). Часть данных этого файла использовалась в главе 15. В качестве зависимых переменных применялись показатели продуктивности запоминания ряда слов сразу после их предъявления: начало1 — для начала ряда; средн1 — для середины ряда; конец1 — для конца ряда. Независимыми переменными являлись: int — интонационное выделение середины ряда (1 — нет, 2 — есть); част — частота встречаемости слов (1 — высокочастотные, 2 — низкочастотные); знач — эмоциональная значимость слов ряда (количественная переменная, используемая в качестве ковариаты). Для примеров этой главы добавим еще три зависимые переменные (начало2, средн2, конец2), соответствующих продуктивности отсроченного (через два дня после предъявления) воспроизведения ряда слов.

Таким образом, с добавлением повторных измерений трех зависимых переменных план исследования усложняется и может быть представлен как комбинация двух внутригрупповых факторов, двух межгрупповых факторов и одной ковариаты. Первый внутригрупповой фактор соответствует части ряда слов и имеет три уровня (начало, середина и конец). Второй внутригрупповой фак-

тор — это отсрочка воспроизведения, имеет два уровня (без отсрочки, с отсрочкой в 2 дня). Нетрудно заметить, что существует шесть комбинаций уровней этих двух факторов, каждой из которых соответствует одна из переменных (табл. 16.1).

**Таблица 16.1.** Сочетания комбинаций части ряда слов и отсрочки их воспроизведения

| **Часть ряда** | **Отсрочка воспроизведения** | |
|---|---|---|
| | Нет | **2 дня** |
| Начало | начало1 | начало2 |
| Середина | средн1 | средн2 |
| Конец | конец1 | конец2 |

Переменные, перечисленные в таблице, по сути представляют собой шесть повторных измерений одной и той же характеристики — продуктивности воспроизведения 8 предъявленных ранее слов. Поэтому для анализа таких данных используется модель многомерного дисперсионного анализа с повторными измерениями[1].

Модель многомерного дисперсионного анализа с повторными измерениями имеет много общего с моделью многомерного дисперсионного анализа без повторных измерений, поэтому материал этой главы в значительной степени опирается на материал предыдущей. В случае если вы пропустили главу 15, настоятельно рекомендуем вернуться к ней, прежде чем приступать к интерпретации результатов приводимых здесь примеров.

# Пошаговые алгоритмы вычислений

При проведении многомерного дисперсионного анализа с повторными измерениями сначала необходимо выполнить три подготовительных шага.

**Шаг 1** Создайте новый файл данных или подготовьте существующий. Все необходимые действия описаны в главе 3.

**Шаг 2** Запустите программу SPSS при помощи значка на рабочем столе или команды Пуск ▸ Программы ▸ SPSS for Windows ▸ SPSS 11.5 for Windows (Start V Programs ▸ SPSS for Windows ▸ SPSS 11.5 for Windows) главного меню Windows. В открывшемся после запуска программы диалоговом окне SPSS for Windows щелкните на кнопке Cancel (Отмена).

[1] Для выполнения многомерного анализа с повторными измерениями требуется полная установка пакета SPSS, включая модули Advanced Models (Дополнительные модели) и Regression Models (Регрессионные модели).

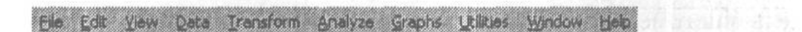
После выполнения этого шага на экране появится окно редактора данных SPSS.



Шаг 3 Откройте файл данных, с которым вы намерены работать (в нашем случае это файл ex02.sav). Если он расположен в текущей папке, то выполните следующие действия:

1. Выберите в меню File (Файл) команду Open ▸ Data (Открытие ▸ Данные) или щелкните на кнопке Open File (Открытие файла) панели инструментов.
2. В открывшемся диалоговом окне дважды щелкните на имени ex02.sav или введите его с клавиатуры и щелкните на кнопке OK.

Независимо от того, открыта программа SPSS только что или какие-то процедуры уже выполнялись, в верхней части главного окна должна присутствовать строка меню (она показана ниже). Пока строка меню присутствует на экране, доступны все команды анализа данных. При этом окно с данными видеть не обязательно.




При работе со сводными таблицами или при редактировании диаграмм некоторые пункты строки меню могут исчезать или меняться. Чтобы вернуться к главному окну и стандартной строке меню, щелкните на кнопке свертывания или восстановления текущего окна.

После завершения шага 3 на экране должно присутствовать окно редактора данных.

**Шаг 4** В меню Analyze (Анализ) выберите команду General Linear Model ▸ Repeated Measures (Общая линейная модель ▸ Повторные измерения). На экране появится диалоговое окно Repeated Measures Define Factor(s) (Повторные измерения: Определение факторов), показанное на рис. 16.1.

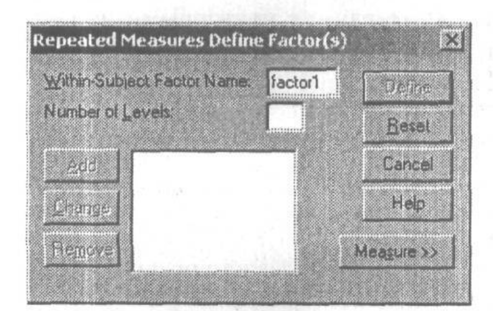



Рис. 16.1. Диалоговое окно Repeated Measures Define Factor(s)

В этом окне вам предстоит задать имена внутригрупповых факторов. Обратите внимание, что имена, которые вы будете вводить в поле Within-Subject Factor Name (Имя внутригруппового фактора), относятся не к существующим переменным из файла данных, а определяют виртуальные переменные для команды

Repeated Measures (Повторные измерения). В нашем примере присвоим этим переменным (внутригрупповым факторам) имена ч_ряда (часть ряда) и отсроч (отсрочка воспроизведения).

Для задания имени внутригрупповой переменной сначала необходимо ввести его в поле Within-Subject Factor Name (Имя внутригруппового фактора), затем задать число уровней переменной в поле Number of Levels (Число уровней) и щелкнуть на кнопке Add (Добавление).

После того как внутригрупповые переменные (факторы) определены, чтобы завершить задание параметров анализа, щелкните на кнопке Define (Определить). На экране появится диалоговое окно Repeated Measures (Повторные измерения), представленное па рис. 16.2.

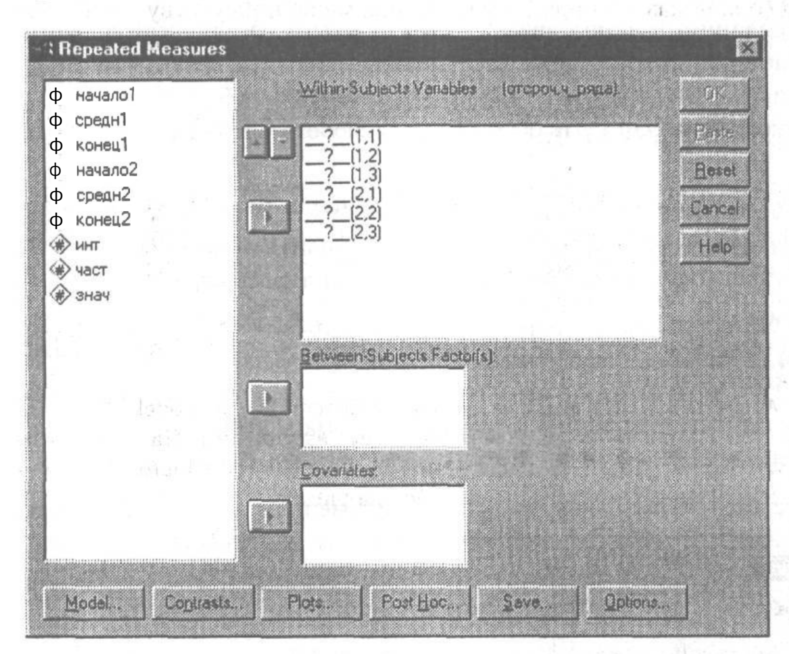



**Рис. 16.2.** Диалоговое окно Repeated Measures

В левой части окна находится список переменных файла ex02.sav. В списке Within-Subjects Variables (Внутригрупповые переменные) вы можете указать соответствие между переменными файла и каждым уровнем любого внутригруппового фактора. В пашем случае в анализ включены два внутригрупповых фактора, один из которых имеет два, а другой — три уровня; таким образом, каждой из шести комбинаций уровней необходимо сопоставить переменную из исходного списка. Пары чисел, указанные в скобках в списке Within-Subjects Variables (Внутригрупповые переменные), определяют, к какой ячейке модели относятся переменные файла, которые далее будут обозначаться как Внутригрупповые переменные (табл. 16.2).

**Таблица 16.2.** Соответствие переменных файла и ячеек модели

| Переменная | Ячейка | Отсрочка | Часть ряда |
|---|---|---|---|
| начало1 | (1,1) | Нет — 1 | Начало — 1 |
| средн1 | (1,2) | Нет — 1 | Середина — 2 |
| конец1 | (1,3) | Нет — 1 | Конец — 3 |
| начало2 | (2,1) | Есть — 2 | Начало — 1 |
| средн2 | (2,2) | Есть — 2 | Середина — 2 |
| конец2 | (2,3) | Есть — 2 | Конец — 3 |

Обратите внимание, что при движении от начала к концу списка переменных файла (первый столбец табл. 16.2) градации фактора «отсрочка» меняются реже, чем градации фактора «часть ряда». Поэтому в качестве первого фактора лучше выбирать тот, который имеет меньше уровней. Несмотря на то что эта иллюстрация относится к случаю двух факторов с небольшим числом уровней, она с легкостью может быть обобщена для любого числа факторов и их уровней.



Для того чтобы ввести какую-либо переменную в список Within-Subjects Variables (Внутригрупповые переменные), достаточно выделить ее имя в списке слева и щелкнуть на верхней кнопке с направленной вправо стрелкой.

По умолчанию SPSS сопоставляет первую введенную переменную первой ячейке модели (в нашем случае — ячейке 1,1), вторую переменную — второй ячейке (1,2), и т. д. Если же вы хотите нарушить данный порядок, воспользуйтесь кнопками с направленными вверх и вниз стрелками, чтобы переместить введенную переменную в нужное место списка ячеек.

После того как все внутригрупповые зависимые переменные анализа определены, вы можете указать все межгрупповые переменные (факторы) и ковариаты. Для этого переместите нужные переменные из исходного списка в списки Between-Subjects Factor(s) (Межгрупповые факторы) и Covariates (Ковариаты).

В нижней части диалогового окна Repeated Measures (Повторные измерения) расположены шесть кнопок: Model (Модель), Contrasts (Контрасты), Plots (Диаграммы), Post Hoc (Постфактум), Save (Сохранение) и Options (Параметры). Щелчок на кнопке Options (Параметры) открывает диалоговое окно, идентичное окну Multivariate: Options (Многомерный анализ: Параметры), с которым мы познакомились в главе 15. Кнопка Plots (Диаграммы) также напоминает свой аналог из глав 14 и 15, но позволяет задавать диаграммы не только для межгрупповых, но и для внутригрупповых факторов. Функции кнопки Post Hoc (Постфактум) также остались прежними; единственным отличием является то, что парные сравнения постфактум применимы только к межгрупповым факторам.

За описанием кнопок Contrasts (Контрасты) и Save (Сохранение) обратитесь к руководству пользователя.

При щелчке на кнопке Model (Модель) открывается диалоговое окно Repeated Measures: Model (Повторные измерения: Модель), представленное на рис. 16.3. В большинстве случаев исследователи предпочитают пользоваться полнофакторной моделью и не меняют положение переключателей в группе Specify Model (Выбор модели). Тем не менее иногда требуется задать набор главных эффектов и взаимодействий вручную. Для этого следует установить переключатель Custom (Настройка), а затем переместить желаемые эффекты и взаимодействия из левых списков в правые. Заметьте, что вы не сможете переместить межгрупповые факторы во внутригрупповую секцию модели и наоборот, так как это некорректно. После того как желаемая модель определена, щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Repeated Measures (Повторные измерения).

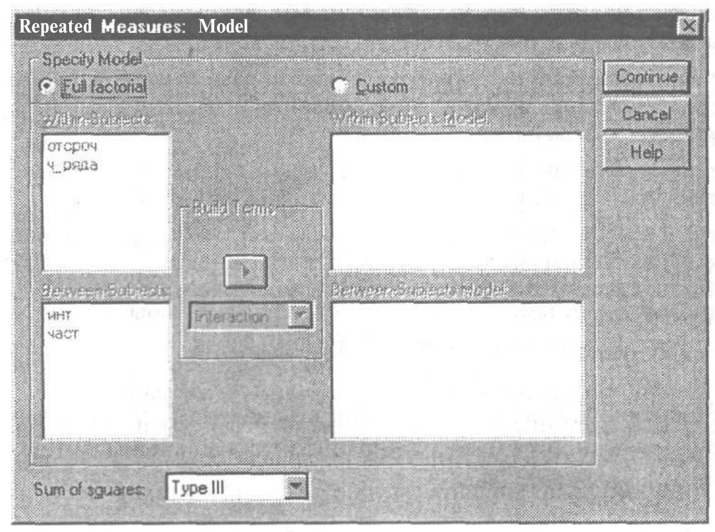



**Рис. 16.3.** Диалоговое окно Repeated Measures: Model

Далее будут приведены два примера использования команды Repeated Measures (Повторные измерения). В первом выполняется многомерный дисперсионный анализ только с повторными измерениями, второй включает не только внутригрупповые, но и межгрупповые факторы, а также ковариату.

В первом примере мы проведем дисперсионный анализ с повторными измерениями, с двумя внутригрупповыми факторами: части ряда (3 градации) и отсрочкой воспроизведения (две градации). С каждой из шести ячеек модели будет связана одна из переменных начало1, средн1, конец1, начало2, средн2, конец2.

Шаг 5 1 После выполнения предыдущего шага у вас должно быть открыто диалоговое окно Repeated Measures Define Factor(s) (Повторные измерения: Определение факторов), показанное на рис. 16.1. Если вы уже успели поработать с этим окном, щелкните на кнопке Reset (Сброс).

1. С помощью клавиши Delete или Backspace удалите имя factorl из поля Within-Subject Factor Name (Имя внутригруппового фактора), введите туда имя отсроч, нажмите клавишу Tab, чтобы переместить фокус ввода в поле Number of Levels (Число уровней), введите число 2 и щелкните на кнопке Add (Добавление).
2. В поле Within-Subject Factor Name (Имя внутригруппового фактора) введите имя ч_ряда, нажмите клавишу Tab , чтобы переместить фокус ввода в поле Number of Levels (Число уровней), введите число 3 и щелкните сначала на кнопке Add (Добавление), а потом — на кнопке Define (Определить), чтобы открыть диалоговое окно Repeated Measures (Повторные измерения), показанное на рис. 16.2.
3. Наведите указатель мыши на переменную начало1, нажмите кнопку мыши и, не отпуская ее, переместите указатель на переменную конец2 и отпустите кнопку. В результате окажутся выделенными шесть переменных: начало1, средн1, конец1, начало2, средн2, конец2.
4. Щелкните на верхней кнопке с направленной вправо стрелкой, чтобы переместить выделенные переменные в список Within-Subjects Variables (Внутригрупповые переменные).
5. Щелкните на кнопке Options (Параметры), чтобы открыть диалоговое окно Repeated Measures: Options (Повторные измерения: Параметры).
6. Установите флажки Descriptive statistics (Описательные статистики) и Estimates of effect size (Оценивать величину эффекта), а затем щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Repeated Measures (Повторные измерения).
7. Щелкните на кнопке OK, чтобы открыть окно вывода.

Следующий пример относится к проведению смешанного анализа с участием одного внутригрупгювого (ч_ряда) и двух межгрупповых факторов (инт, част). Кроме того, в анализе участвует ковариата знач.

Шаг 5а После выполнения шага 4 у вас должно быть открыто диалоговое окно Repeated Measures Define Factor(s) (Повторные измерения: Определение факторов), показанное на рис. 16.1. Если вы уже успели поработать с этим окном, щелкните на кнопке Reset (Сброс).

1. С помощью клавиши Delete или Backspace удалите имя factorl из поля Within-Subject Factor Name (Имя внутри.группового фактора), введите туда имя ч_ряда, нажмите клавишу Tab, чтобы переместить фокус ввода в поле Number of Levels (Число уровней), введите число 3 и щелкните на кнопке Add (Добавление).
2. Наведите указатель мыши на переменную начало2, нажмите кнопку мыши и, не отпуская ее, переместите указатель на переменную конец2 и отпустите кнопку. В результате окажутся выделенными три переменные: начало2, средн2 и конец2.

3. Щелкните па верхней кнопке с направленной вправо стрелкой, чтобы переместить выделенные переменные в список Within-Subjects Variables (Внутригрупповые переменные).
4. Щелкните сначала на переменной инт, чтобы выделить ее, а затем — па средней кнопке с направленной вправо стрелкой, чтобы переместить переменную в список Between-Subjects Factor(s) (Межгрупповые факторы).
5. Щелкните сначала на переменной знач, чтобы выделить ее, а затем — на нижней кнопке с направленной вправо стрелкой, чтобы переместить переменную в список Covariates (Ковариаты).
6. Щелкните на кнопке Options (Параметры), чтобы открыть диалоговое окно Repeated Measures: Options (Повторные измерения: Параметры).
7. Установите флажки Descriptive statistics (Описательная статистика) и Estimates of effect size (Оценивать величину эффекта), а затем щелкните па кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Repeated Measures (Повторные измерения).
8. Щелкните па кнопке OK, чтобы открыть окно вывода.

После выполнения шага 5 программа автоматически активизирует окно вывода. Для просмотра результатов вы при необходимости можете воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

# Печать результатов и выход из программы

Ниже описана типичная процедура печати результата статистического анализа (или нескольких анализов). После выполнения шага 5 должно быть открыто окно вывода.

**Шаг 6 III** В окне вывода укажите фрагменты, выводимые на печать (см. раздел «Окно вывода» в главе 2), в меню File (Файл) выберите команду Print (Печать), при необходимости задайте параметры печати и щелкните на кнопке OK.

Последнее, что необходимо сделать после завершения исследования и печати результатов, — это выйти из программы SPSS.

**Шаг 7** Для выхода из программы в меню File (Файл) выберите команду Exit (Выход).

Иногда после выполнения команды Exit (Выход) на экране могут появляться небольшие диалоговые окна с вопросом о необходимости сохранения сделанных в файлах изменений и кнопками, описывающими возможные варианты ответа. Для завершения работы просто щелкайте на соответствующих кнопках.

# Представление результатов

Большая часть результатов, генерируемых командой Repeated Measures (Повторные измерения), аналогична результатам многомерного дисперсионного анализа (MANOVA), поэтому ниже будут представлены лишь те фрагменты, которые являются специфичными для внутригрупповых моделей. Также не приводятся некоторые результаты, которые являются избыточными и практически не требуют интерпретации, например внутригрупповые контрасты (Tests of Within-Subjects Contrasts).

## Многомерные критерии

Таблица, показанная на рис. 16.4, получена после выполнения шага 5.

**Multivariate Tests**

| Effect | | Value | F | Hypothesis df | Error df | Sig. |
|---|---|---|---|---|---|---|
| ОТСРОЧ | Pillai's Trace | .862 | 119 | 1.000 | 19.0 | .000 |
| | Wilks' Lambda | .138 | 119 | 1.000 | 19.0 | .000 |
| | Hotelling's Trace | 6.255 | 119 | 1.000 | 19.0 | .000 |
| | Roy's Largest Root | 6.255 | 119 | 1.000 | 19.0 | .000 |
| Ч_РЯДА | Pillai's Trace | .295 | 3.8 | 2.000 | 18.0 | .043 |
| | Wilks' Lambda | .705 | 3.8 | 2.000 | 18.0 | .043 |
| | Hotelling's Trace | .418 | 3.8 | 2.000 | 18.0 | .043 |
| | Roy's Largest Root | .418 | 3.8 | 2.000 | 18.0 | .043 |
| ОТСРОЧ * Ч_РЯДА | Pillai's Trace | .057 | .542 | 2.000 | 18.0 | .591 |
| | Wilks' Lambda | .943 | .542 | 2.000 | 18.0 | .591 |
| | Hotelling's Trace | .060 | .542 | 2.000 | 18.0 | .591 |
| | Roy's Largest Root | .060 | .542 | 2.000 | 18.0 | .591 |

**Рис. 16.4.** Фрагмент окна вывода после выполнения шага 5 (многомерные критерии)

Как видно из таблицы, главные эффекты факторов отсрочки воспроизведения и части ряда являются значимыми ($p < 0{,}05$), поэтому для их интерпретации можно непосредственно сравнить средние значения соответствующих ячеек либо применить $t$-критерий. В данном случае мы можем сделать вывод, что с отсрочкой ухудшается воспроизведение предъявленных слов; эффективность воспроизведения начала ряда хуже, чем середины и конца ряда; эти эффекты проявляются независимо друг от друга. Все это иллюстрирует диаграмма (рис. 16.5), которую вы можете легко построить, щелкнув на кнопке Plots (Диаграммы) в окне Repeated Measures (Повторные измерения), показанном ранее на рис. 16.2 (о функции этой кнопки см. главу 14, о работе с диаграммами см. главу 5).

В случае если вы используете смешанную модель, то в выводимые данные будут включены эффекты взаимодействий для внутригрупповых и межгрупповых факторов. Если в модели присутствуют ковариаты, то выводимые данные будут

содержать результаты **их** взаимодействия с внутригрупповыми факторами. Ниже дана трактовка терминов, используемых программой в окне вывода и относящихся к многомерным критериям.

- Value (Значение) — в этом столбце представлены значения многомерных критериев, определяющих достоверность различия повторных измерений для соответствующих внутригрупповых факторов.
- F (*F*-критерий) — значение *F*-критерия.
- Hypothesis df (Гипотетическое число степеней свободы) — произведение числа уровней факторов, уменьшенных на единицу.
- Error df (Ошибка числа степеней свободы) — величина, вычисляемая различными способами в зависимости от теста.
- Sig. (Значимость) — уровень значимости для соответствующего *F*-критерия.

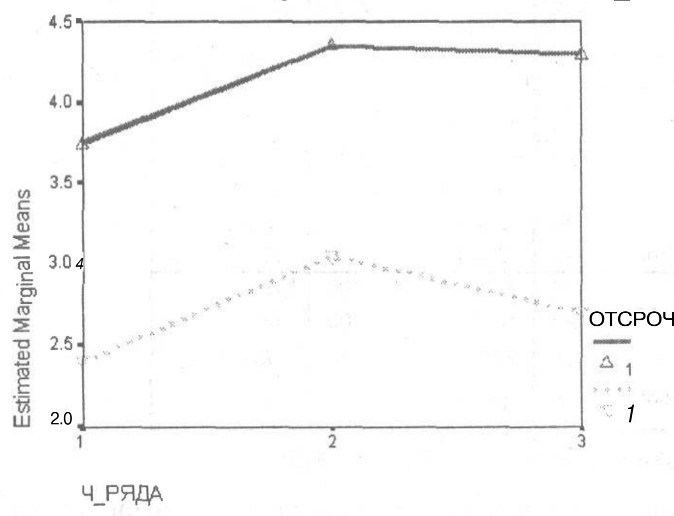



**Рис. 16.5.** Фрагмент окна вывода после выполнения шага 5 (график средних)

## Критерии для внутригрупповых эффектов

Таблица, показанная на рис. 16.6, получена после выполнения шага 5а.

Полученные результаты во многом дублируют результаты применения многомерных критериев. Таблица содержит главный эффект внутригруппового фактора (ч_ряда), а также эффекты его взаимодействия с ковариатой (знач) и межгрупповым фактором (инт). Все эффекты оказываются статистически достоверными. Столбец Partial Eta Squared (Частичный квадрат Эта) иллюстрирует меру эффекта ($\eta^2$), показывающую процент дисперсии, обусловленный воздействием фактора или взаимодействием факторов. Для интерпретации статистически достоверных

результатов можно обратиться к описательным статистикам (средним) или к графикам средних значений. Для данного случая график средних значений приведен на рис. 16.7. Этот график можно построить, щелкнув на кнопке Plots (Диаграммы) в окне Repeated Measures (Повторные измерения), показанном ранее на рис. 16.2 (о функции этой кнопки см. главу 14, о работе с диаграммами см. главу 5).

**Tests of Within-Subjects Effects**

Measure: MEASURE_1

| Source | | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|---|---|---|---|---|---|---|---|
| Ч_РЯДА | Sphericity Assumed | 4.816 | 2 | 2.408 | 12.2 | .000 | .417 |
| | Greenhouse-Geisser | 4.816 | 1.80 | 2.677 | 12.2 | .000 | .417 |
| | Huynh-Feldt | 4.816 | 2.00 | 2.408 | 12.2 | .000 | .417 |
| | Lower-bound | 4.816 | 1.00 | 4.816 | 12.2 | .003 | .417 |
| Ч_РЯДА * ЗНАЧ | Sphericity Assumed | 2.869 | 2 | 1.435 | 7.247 | .002 | .299 |
| | Greenhouse-Geisser | 2.869 | 1.80 | 1.595 | 7.247 | .003 | .299 |
| | Huynh-Feldt | 2.869 | 2.00 | 1.435 | 7.247 | .002 | .299 |
| | Lower-bound | 2.869 | 1.00 | 2.869 | 7.247 | .015 | .299 |
| Ч_РЯДА * ИНТ | Sphericity Assumed | 4.562 | 2 | 2.281 | 11.5 | .000 | .404 |
| | Greenhouse-Geisser | 4.562 | 1.80 | 2.536 | 11.5 | .000 | .404 |
| | Huynh-Feldt | 4.562 | 2.00 | 2.281 | 11.5 | .000 | .404 |
| | Lower-bound | 4.562 | 1.00 | 4.562 | 11.5 | .003 | .404 |
| Error(Ч_РЯДА) | Sphericity Assumed | 6.731 | 34 | .198 | | | |
| | Greenhouse-Geisser | 6.731 | 30.6 | .220 | | | |
| | Huynh-Feldt | 6.731 | 34.0 | .198 | | | |
| | Lower-bound | 6.731 | 17.0 | .396 | | | |

Рис. 16.6. Фрагмент окна вывода после выполнения шага 5а (критерии для внутригрупповых эффектов)

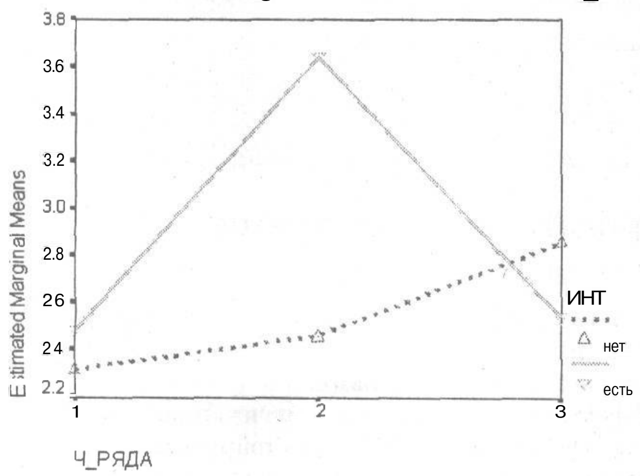



**Рис.** 16.7. Фрагмент окна вывода после выполнения шага 5а (график средних)

Судя по графику, взаимодействие факторов инт и ч_ряда обусловлено более продуктивным воспроизведением слов середины ряда в случае их интонационного выделения. Главный эффект фактора ч_ряда обусловлен тем, что слова в начале ряда воспроизводились хуже, чем слова в середине и конце ряда.

## Критерии для межгрупповых эффектов

Таблица, показанная на рис. 16.8, получена после выполнения шага 5а.

**Tests of Between-Subjects Effects**

Measure: MEASURE_1

Transformed Variable: Average

| Source | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|---|---|---|---|---|---|---|
| Intercept | 5.243 | 1 | 5.243 | 6.891 | .018 | .288 |
| ЗНАЧ | 22.564 | 1 | 22.564 | 28.7 | .000 | .636 |
| ИНТ | 1.383 | 1 | 1.383 | 1.818 | .195 | .097 |
| Error | 12.936 | 17 | .761 | | | |

**Рис. 16.8.** Фрагмент окна вывода после выполнения шага 5а (критерии для межгрупповых эффектов)

Межгрупповые эффекты в результатах выводятся вместе с эффектами ковариат. Большая часть информации, содержащейся в этой таблице, повторяет предыдущие результаты.

# 17 Простая линейная регрессия



Команды подменю Regression (Регрессия) позволяют выполнять как простой, так и множественный регрессионный анализ. Простой регрессионный анализ обсуждается в этой главе, а множественный — в следующей. Такое разбиение одного статистического раздела было сделано нами в целях упрощения изложения материала, касающегося основ регрессионного анализа. Если вы не знакомы с множественным регрессионным анализом, то эта глава послужит вам введением. Мы рассмотрим такие понятия, как прогнозируемые значения зависимой переменной и уравнение регрессии, покажем связь между простой регрессией и корреляцией двух переменных, рассмотрим влияние одной переменной на дисперсию другой, а также ознакомимся с оценкой криволинейности связи двух переменных.

## Простая регрессия

Зачастую, имея информацию об одной характеристике, мы пытаемся на основе этой информации делать выводы о другой характеристике, связанной с исходной. Например, если мы знаем рост человека, мы могли бы попытаться оценить его возможный вес. Например, зная, что человек имеет рост 214 см, с определенной вероятностью мы можем прогнозировать, что вес этого человека превысит 91 кг. Можно привести множество пар связанных между собой величин: коэффициент интеллекта (IQ) и академическая успеваемость, число сокращений мышц ног в секунду и скорость бега, калорийность пищи и вес, симпатия к человеку и желание оказать ему помощь и т. д. Каждый день мы оцениваем связи различных факторов между собой. Именно на проведение таких оценок и нацелена простая регрессия. Простая регрессия не может дать абсолютно достоверного результата, однако способна ответить на вопрос о *связи* неременных, а также по заданному значению одной переменной рассчитать наиболее вероятное значение другой переменной.

Для иллюстрации практического применения простой регрессии воспользуемся новым файлом данных с именем exam.sav. Он содержит 10-балльную оценку нервной возбудимости (тревожности) 36 студентов и количество решенных ими зачетных тестовых задний (из 20 возможных). Гипотеза о линейности отношения этих двух переменных говорит о том, что чем выше нервная возбудимость студента, тем выше его результативность (например, потому, что спокойных студентов меньше волнуют их знания, а тревожные студенты проводят больше времени за подготовкой к зачету). В качестве зависимой переменной (критерия) выступает переменная тест, а в качестве независимой переменной (предиктора) - переменная трев. Мы будем пытаться прогнозировать значения переменной тест по известным значениям переменной трев, используя ***уравнение регрессии.*** Уравнение регрессии формируется на основе общего уравнения, связывающего фактическую успеваемость студента и нервную возбудимость:

$$тест_{истина} = \textit{константа} + \textit{коэффициент} \times трев + \textit{остаток}.$$

Здесь $тест_{истина}$ — переменная, отражающая реальный результат выполнения тестового задания, ***константа*** — некоторая константа, ***коэффициент*** — регрессионный коэффициент при значении оценки тревожности, ***остаток*** — остаток. Для нашего примера константа равна 9,3114, а коэффициент регрессии — 0,6751. Соответственно, уравнение для прогноза результата зачетного тестирования выглядит следующим образом:

$$тест_{прогноз} = 9{,}3114 + 0{,}6751 \times трев,$$

$$тест_{истина} - тест_{прогноз} + \textit{остаток}.$$

Прогнозируемое значение будет отличаться от истинного значения. Разница отражает тот факт, что рассчитываемые результаты экзамена никогда не бывают абсолютно точными, и чтобы получить истинный результат, необходимо ввести в уравнение член, равный разности прогнозируемого и реального значений. Этот член и называют ***остатком:***

*реальное значение - прогнозируемое значение + остаток.*

Понятие остатка не представляет сложности для понимания, и мы еще вернемся к остаткам в последней главе книги.

Можно предположить, что помимо нервной возбудимости на результаты зачетного тестирования влияют другие факторы. Регрессионный анализ, учитывающий влияние нескольких факторов, называется ***множественным*** и рассматривается в следующей главе.

Далее перечислены величины, которые вычисляются при проведении регрессионного анализа с помощью команд подменю Regression (Регрессия).

- Коэффициент, характеризующий связь между значениями зависимой и независимой переменных, обозначается прописной буквой $R$, набранной курсивом, и представляет собой не что иное, как знакомый нам по предыдущим главам коэффициент корреляции. Использование прописной буквы вместо

строчной в выводимых данных регрессионного анализа объясняется тем, что в общем случае при выполнении анализа вычисляются *множественные* корреляции с участием нескольких независимых переменных. Корреляция подробно рассмотрена в главе 9.

- Для рассчитанного значения множественного коэффициента корреляции $R$ программа определяет уровень значимости $p$. Как и ранее, величина $p < 0{,}05$ свидетельствует о значимой корреляции переменных. При $p > 0{,}05$ вероятность случайности результата считается слишком высокой, и в этом случае говорят, что связь между переменными слабая или не обнаружена.
- Значение $R^2$ является обычным квадратом величины $R$, однако этот коэффициент имеет собственный смысл. Коэффициент $R^2$ характеризует долю дисперсии одной переменной, обусловленной воздействием другой переменной. Так, для переменных трев и тест значение $R = 0{,}546$, а $R^2 = 0{,}298$. Это означает, что 29,8 % дисперсии переменной тест объясняется влиянием независимой переменной трев.
- Константу и коэффициент уравнения регрессии часто называют $B$-величинами. $B$-величины характеризуют связь между значениями переменных трев и тест,

# Оценка криволинейности

Выдвинутая в предыдущем разделе гипотеза о том, что увеличение нервной возбудимости перед экзаменом всегда улучшает результат студента, очевидно, вызывает сомнения. Более разумными кажутся следующие соображения. При низкой возбудимости результаты экзамена должны быть низкими, поскольку излишнее спокойствие снижает потребность студента в подготовке. С увеличением возбудимости результат до определенного момента должен улучшаться, однако слишком возбудимые студенты вряд ли способны сконцентрироваться и показать хороший результат. Поэтому наилучших показателей должны добиваться те студенты, чей уровень возбудимости является промежуточным. При регрессионном анализе, вне зависимости от того, является он множественным или простым, измеряется линейное соотношение между зависимой и независимой переменными. В приведенном примере мы видим значительную корреляцию между переменными трев и тест ($R$ - 0,546, $p < 0{,}001$), однако возможная ошибка прогноза велика (только 29,8 % дисперсии переменной тест объясняется влиянием переменной трев). Можно предположить, что если изменить вид общего уравнения (например, включить в него квадрат переменной трев), то прогнозируемые значения будут ближе к реальным.

На рис. 17.1 наглядно демонстрируется характер отношений между переменными трев и тест.

По вертикальной оси диаграммы отложены значения переменной тест, по горизонтальной — значения переменной трев. Очевидно, что зависимость между

переменными не является линейной: у нее имеется выраженный максимум, а при движении к краям диаграммы наблюдается убывание значений переменной тест. Нельзя не отметить, что только путем графической интерпретации невозможно достоверно установить характер отношений между переменными. Необходимым условием является применение статистических критериев.

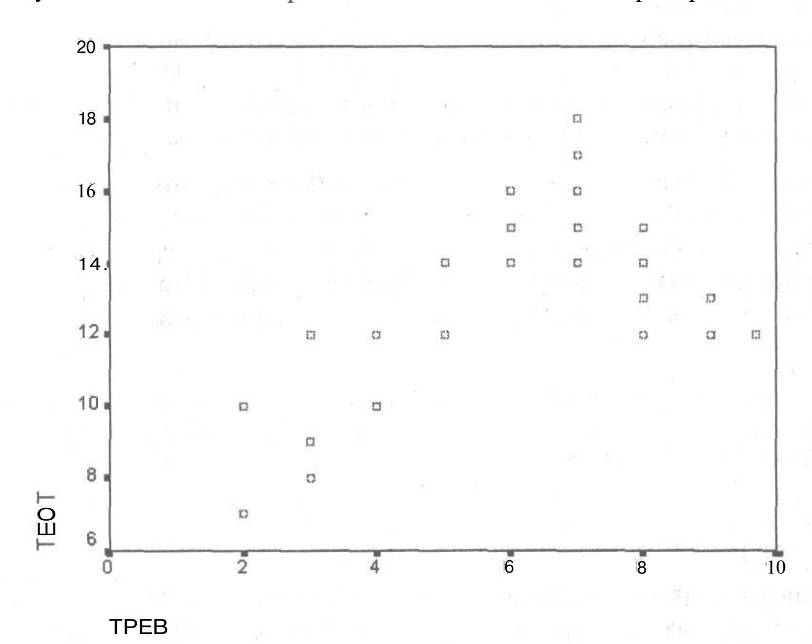



**Рис. 17.1.** Диаграмма рассеивания, демонстрирующая криволинейную зависимость

Для того чтобы статистически оценить криволинейность, в подменю Regression (Регрессия) предусмотрена команда Curve Estimation (Оценка криволинейности). В ее диалоговом окне необходимо задать зависимую переменную (тест), независимую переменную (трев) и установить флажки Linear (Линейная зависимость) и Quadratic (Квадратичная зависимость). После щелчка на кнопке ОК появятся результаты, приведенные ниже, а также диаграмма, показанная на рис. 17.2.

```
Depend. Mth Rsq d.f. F Sigf b0 b1 b2
ТЕСТ LIN .298 34 14.47 .001 9.3114 .6751
ТЕСТ QUA .675 33 34.28 .000 .1615 4.4896 -.3381
```

По диаграмме можно определить, насколько близка к линейной или квадратичной зависимость между значениями нeременных. В результаты включены значения коэффициентов *B* регрессии (b0, Ы, Ь2), поэтому несложно составить линейное и квадратичное уравнения регрессии для прогнозируемых значений. На диаграмму помимо линейной и квадратичной зависимостей нанесен разброс данных файла. Обратите внимание па сходство диаграмм, представленных на рис. 17.1 и 17.2, а также на то, что константа и коэффициенты уравнений указаны в первых двух строках выводимых результатов.

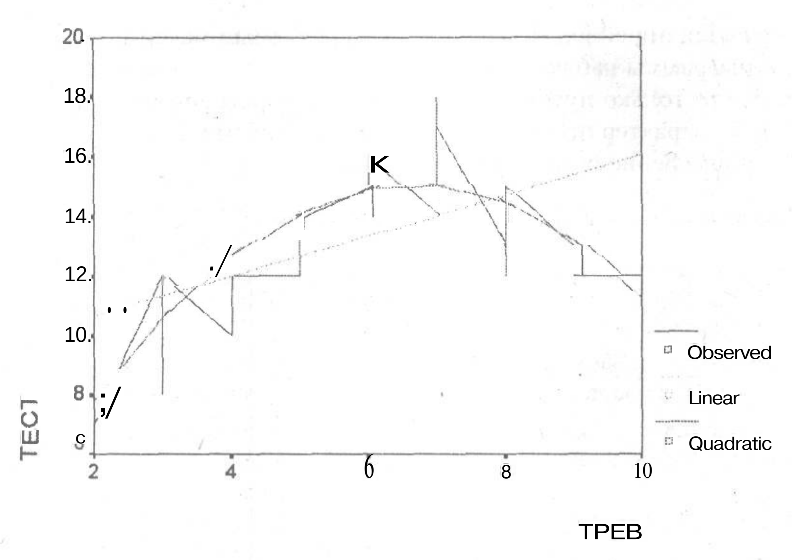



**Рис. 17.2.** Диаграмма, демонстрирующая линейную и криволинейную тенденции

Для линейной регрессии уравнение имеет вид:

$$тест_{прогноз} = 9{,}3114 + 0{,}6751(трев).$$

А уравнение для квадратичной регрессии выглядит следующим образом:

$$тест_{прогноз} - 0{,}1615 + 4{,}4896(трев) - 0{,}3381(трев)^2.$$

Обратите внимание, что в случае линейной регрессии величина $R^2$ (столбец Rsq в таблице выводимых результатов) равна 0,298, то есть 29,8 % дисперсии переменной тест обусловлено воздействием со стороны переменной трев. В то же время для квадратичной регрессии, которая учитывает и линейную, и криволинейную связи, $R^2$ = 0,675. Таким образом, переменная трев и ее квадрат обусловливают 67,5 % дисперсии переменной тест. Малый $p$-уровень для обоих уравнений свидетельствует об очень высокой статистической достоверности полученных результатов. Очевидно, что квадратичная регрессия описывает отношения между переменными тест и трев более адекватно, чем линейная.

# Пошаговые алгоритмы вычислений

Для проведения регрессионного анализа сначала необходимо выполнить три подготовительных шага (обратите внимание, что в данном случае мы работаем с файлом exam.sav).

**Шаг 1** Создайте новый файл данных или подготовьте существующий. Все необходимые действия описаны в главе 3.

**Шаг 2** Запустите программу SPSS при помощи значка на рабочем столе или команды Пуск ▸ Программы ▸ SPSS for Windows ▸ SPSS 11.5 for Windows (Start > Programs ▸ SPSS for Windows ▸ SPSS 11.5 for Windows) главного меню Windows. В открывшемся после запуска программы диалоговом окне SPSS for Windows щелкните на кнопке Cancel (Отмена).

После выполнения этого шага на экране появится окно редактора данных SPSS.

**Шаг 3** Откройте файл данных, с которым вы намерены работать (в нашем случае это файл exam.sav). Если он расположен в текущей папке, то выполните следующие действия:

1. Выберите в меню File (Файл) команду Open ▸ Data (Открытие ▸ Данные) или щелкните на кнопке Open File (Открытие файла) панели инструментов.
2. В открывшемся диалоговом окне дважды щелкните на имени exam.sav или введите его с клавиатуры и щелкните на кнопке OK.

Независимо от того, открыта программа SPSS только что или какие-то процедуры уже выполнялись, в верхней части главного окна должна присутствовать строка меню (она показана ниже). Пока строка меню присутствует на экране, доступны все команды анализа данных. При этом окно с данными видеть не обязательно.




**При** работе со сводными таблицами или при редактировании диаграмм некоторые пункты строки меню могут исчезать или меняться. Чтобы вернуться к главному окну и стандартной строке меню, щелкните на кнопке свертывания или восстановления текущего окна.

## Простой регрессионный анализ

После завершения шага 3 на экране должно присутствовать окно редактора данных.

**Шаг 4** В меню Analyze (Анализ) выберите команду Regression ▸ Linear (Регрессия ▸ Линейная регрессия). На экране появится диалоговое окно Linear Regression (Линейная регрессия), показанное на рис. 17.3.

Диалоговое окно Linear Regression (Линейная регрессия) позволяет проводить разные виды регрессионного анализа. Поскольку в данном случае выполняется простой регрессионный анализ, некоторые из элементов интерфейса этого окна мы рассматривать не будем.

В списке, находящемся в левой части диалогового окна, содержатся имена лишь двух переменных файла exam.sav — это имена трев и тест. Позже нами будет создана еще одна переменная с именем трев2. Для того чтобы провести регрессионный анализ, достаточно выполнить несколько простых действий: поместить переменную тест в ноле Dependent (Зависимая переменная), переменную трев — в список

Independent(s) (Независимые переменные) и щелкнуть на кнопке OK. В результате программа вычислит величины *R*, $R^2$, *F*, соответствующие им *p*-уровни, *B*-величины (коэффициенты и константы уравнения регрессии), а также β-величины, характеризующие степень зависимости между значениями зависимой и независимой переменных.

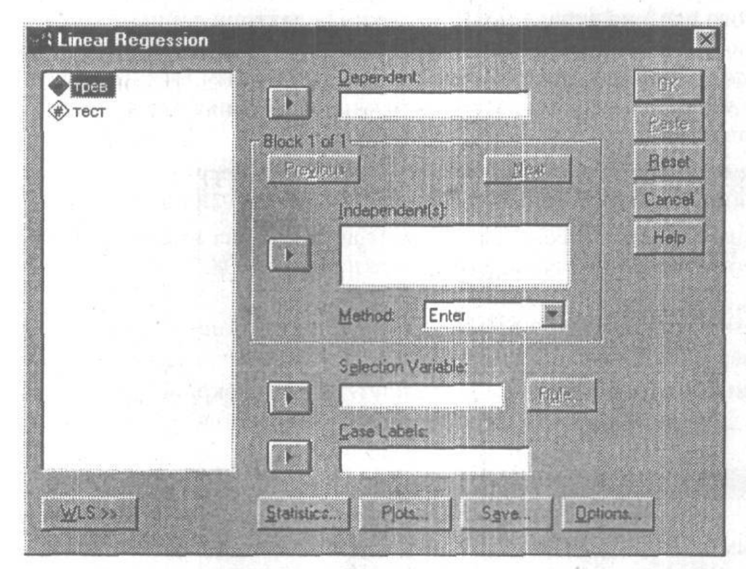



**Рис. 17.3.** Диалоговое окно Linear Regression

В следующем примере мы приводим последовательность действий, необходимую для выполнения простого регрессионного анализа с участием независимой переменной трев и зависимой переменной тест.

**Шаг 5** После выполнения шага 4 должно быть открыто диалоговое окно Linear Regression (Линейная регрессия), показанное па рис. 17.3.

1. Щелкните сначала на переменной тест, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в поле Dependent (Зависимая переменная).
2. Щелкните сначала на переменной трев, чтобы выделить ее, а затем — на второй сверху кнопке со стрелкой, чтобы переместить переменную в список Independent(s) (Независимые переменные).
3. Щелкните на кнопке OK, чтобы открыть окно вывода.

## Анализ криволинейных зависимостей

Мы рассмотрим две пошаговые процедуры, относящиеся к анализу криволинейных зависимостей. Первая позволяет получить результаты, представленные в начале этой главы (таблицу из двух строк и диаграмму, показанную на рис. 17.2), вторая приводит к созданию специальной «квадратичной» переменной, которую при

желании можно использовать в дальнейших вычислениях. Несмотря на то что криволинейные зависимости рассматриваются в контексте простого регрессионного анализа, все сказанное в этой главе справедливо и для множественного регрессионного анализа.

После завершения шага 3 па экране должно присутствовать окно редактора данных.

**Шаг 4а** В меню Analyze (Анализ) выберите команду Regression ▸ Curve Estimation (Регрессия ▸ Оценка криволинейности). На экране появится диалоговое окно Curve Estimation (Оценка криволинейности), показанное на рис. 17.4.

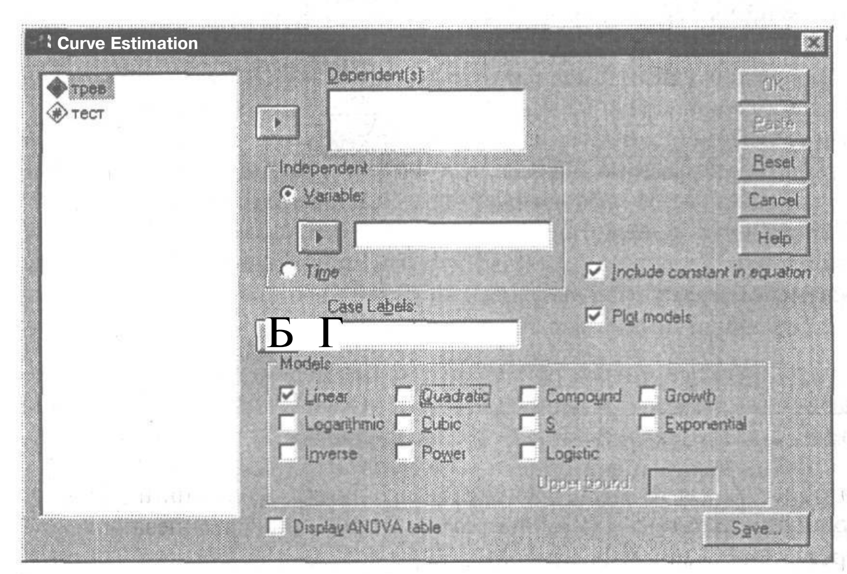



Рис. 17.4. Диалоговое окно Curve Estimation

Первоначальные действия в окне Curve Estimation (Оценка криволинейности) те же, что и в окне Linear Regression (Линейная регрессия): поместите переменную тест в список Dependent(s) (Зависимые переменные), а переменную трев — в поле Independent (Независимая переменная). Обратите внимание на то, что по умолчанию в диалоговом окне установлены три флажка: флажок Include constant in equation (Включить константу в уравнение), а также флажок Plot Models (Модели диаграмм), позволяющий создать диаграмму (см. рис. 17.2), и флажок Linear (Линейная зависимость), обеспечивающий проверку прямолинейности зависимости и включение в диаграмму прямой линии.

Как можно видеть, в группе Models (Модели) помимо флажка Linear (Линейная зависимость) имеется множество других флажков, позволяющих проводить проверки на соответствие зависимости между переменными по самым разным видам кривых и включать эти кривые в диаграмму. В следующем примере мы продемонстрируем проверку зависимости на квадратичность (Quadratic).

**Шаг 5а** Ниисс перечислены действия, позволяющие провести проверку криволинейности. После выполнения шага 4а должно быть открыто диалоговое окно Curve Estimation (Оценка криволинейности), показанное на рис. 17.4.

1. Щелкните сначала па переменной тест, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в список Dependent(s) (Зависимые переменные).
2. Щелкните сначала на переменной трев, чтобы выделить се, затем — на второй сверху кнопке со стрелкой, чтобы переместить переменную в поле Independent (Независимая переменная); установите флажок Quadratic (Квадратичная зависимость).
3. Щелкните на кнопке OK, чтобы открыть окно вывода.

При составлении квадратного уравнения регрессии SPSS создает новую переменную, значения которой равны квадратам соответствующих значений переменной трев. Как вы, наверное, помните из школьного курса алгебры, квадратичная зависимость графически имеет вид параболы — кривой с двумя ветвями, которые могут быть направлены вверх, если коэффициент при квадратном члене положительный, или вниз, если коэффициент при квадратном члене отрицательный. Поскольку квадратное уравнение имеет не только квадратный, по и линейный член, изображаемый в виде наклонной прямой, конец левой ветви параболы расположен ниже, чем конец правой ветви.

Для того чтобы включить в анализ регрессии квадрат переменной трев, нам необходимо «вручную» создать переменную трев2, содержащую квадраты значений переменной трев, а затем указать ее в качестве независимой переменной. Процесс создания новых неременных описан в разделе «Преобразование данных» главы 4. Здесь мы приведем все необходимые инструкции, однако для более детального изучения вопроса вам следует обратиться к упомянутому разделу, в котором, в частности, рассказывается, что создание новой переменной начинается с диалогового окна Compute Variable (Вычисление переменной).

**Шаг 5б** На этом шаге мы создадим новую переменную трев2 и включим ее в последующий регрессионный анализ в качестве независимой. Перед выполнением следующих инструкций у вас должно быть открыто окно редактора данных.

1. В меню Transform (Преобразование) выберите команду Compute (Вычислить). На экране появится диалоговое окно Compute Variable (Вычисление переменной).
2. В поле Target Variable (Целевая переменная) введите имя трев2.
3. Щелкните сначала на переменной трев, чтобы выделить ее, а затем — па кнопке со стрелкой, чтобы ввести переменную в условие отбора, составляемое в поле Numeric Expression (Выражение).
4. Следом за именем переменной в поле Numeric Expression (Выражение) введите две звездочки (**), двойку (2) и щелкните па кнопке OK, чтобы вернуться в окно редактора данных.

5. В меню Analyze (Анализ) выберите команду Regression ▸ Linear (Регрессия ▸ Линейная регрессия), чтобы открыть диалоговое окно Linear Regression (Линейная регрессия), показанное на рис. 17.3. Если вы уже работали с этим окном, щелкните на кнопке Reset (Сброс).
6. Щелкните сначала на переменной тест, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в поле Dependent (Зависимая переменная).
7. Щелкните сначала на переменной трев, чтобы выделить ее, а затем — на второй сверху кнопке со стрелкой, чтобы переместить переменную в список Independent(s) (Независимые переменные).
8. Повторите предыдущее действие для переменной трев2.
9. Щелкните на кнопке OK, чтобы открыть окно вывода.

После выполнения шага 5 программа автоматически активизирует окно вывода. Для просмотра результатов вы при необходимости можете воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

# Печать результатов и выход из программы

Ниже описана типичная процедура печати результатов статистического анализа (или нескольких анализов). После выполнения шага 5 должно быть открыто окно вывода.

**Шаг 6** В окне вывода укажите фрагменты, выводимые па печать (см. раздел «Окно вывода» в главе 2), в меню File (Файл) выберите команду Print (Печать), при необходимости задайте параметры печати и щелкните на кнопке OK.

Последнее, что необходимо сделать после завершения исследования и печати результатов, — это выйти из программы SPSS.

**Шаг 7** Для выхода из программы в меню File (Файл) выберите команду Exit (Выход).

Иногда после выполнения команды Exit (Выход) на экране могут появляться небольшие диалоговые окна с вопросом о необходимости сохранения сделанных в файлах изменений и кнопками, описывающими возможные варианты ответа. Для завершения работы просто щелкайте па соответствующих кнопках.

# Представление результатов

## Простой регрессионный анализ

На рис. 17.5 приведены фрагменты выводимых данных, сгенерированные программой после выполнения шага 5.

**Variables Entered/Removed**[b]

| Model | Variables Entered | Variables Removed | Method |
|---|---|---|---|
| 1 | TPEB[a] | . | Enter |

a. All requested variables entered.

b. Dependent Variable: TECT

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .546[a] | .298 | .278 | 2.30561 |

a. Predictors: (Constant), TPEB

**ANOVA**[b]

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 76.901 | 1 | 76.901 | 14.466 | .001[a] |
| | Residual | 180.738 | 34 | 5.316 | | |
| | Total | 257,639 | 35 | | | |

a. Predictors: (Constant), TPEB

b. Dependent Variable: TECT

**Coefficients**[a]

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 9.311 | 1.118 | | 8.327 | .000 |
| | TPEB | .675 | .177 | .546 | 3.803 | .001 |

a. Dependent Variable: TECT

**Рис. 17.5.** Фрагмент окна вывода после выполнения шага 5

Как можно видеть, между переменными тест и трев имеется значимая линейная связь. Это означает, что с увеличением нервной возбудимости (тревожности) студента количество выполненных им тестовых заданий на зачете также имеет тенденцию к увеличению. Далее дана трактовка терминов, используемых программой в окне вывода.

- R — поскольку в анализе участвовала единственная независимая переменная, эта величина равна коэффициенту корреляции ($r$) между переменными тест и трев.
- R Square (Квадрат R) — квадрат величины $R$ ($R^2$), равный доле дисперсии переменной тест, обусловленной воздействием переменной трев.

- Adjusted R Square (Исправленный квадрат R) — исправленная величина $R^2$. Величина $R^2$, используемая в расчетах, на практике оказывается несколько завышенной. Исправленная величина $R^2$ менее формальна и ближе к реальным результатам.
- Std. Error (Стандартная ошибка) — в таблице Model Summary это стандартное отклонение оценок значений зависимой переменной тест.
- Regression (Регрессия) — статистики, оценивающие долю дисперсии зависимой переменной, обусловленную влиянием независимых переменных.
- Residual (Остаток) — статистики, оценивающие долю дисперсии зависимой переменной, не обусловленную влиянием независимых переменных.
- df (Число степеней свободы) — для регрессии это число равно числу независимых переменных. Для остатка это число равно разности размера выборки и числа степеней свободы регрессии, уменьшенной на единицу (36 - 1 - 1 = 34).
- Sum of Squares (Сумма квадратов) — для регрессии это сумма квадратов между группами; для остатка — сумма квадратов внутри групп.
- Mean Square (Средний квадрат) — отношение суммы квадратов к числу степеней свободы.
- F ($F$-критерий) — отношение среднего квадрата для регрессии к среднему квадрату для остатка.
- Sig. (Значимость) — вероятность случайности полученного результата.
- B — коэффициент и константа линейного уравнения регрессии:

$$\text{тест}_{\text{прогноз}} = 9{,}311 + 0{,}675(\text{трев}).$$

- Std. Error (Стандартная ошибка) — в таблице Coefficients это характеристика стабильности коэффициента *B*, равная стандартному отклонению коэффициентов *B*, рассчитанных для большого числа выборок из генеральной совокупности.
- Beta (Бета) — стандартизованный коэффициент регрессии (р). Представляет собой коэффициент *B* для стандартизованных значений переменной трев. Для линейных отношений эта величина всегда лежит в диапазоне от –1,0 до 1,0, а для криволинейных отношений может выходить за границы этого диапазона.
- t — отношение коэффициента *B* к его стандартной ошибке.

## Регрессионный анализ криволинейной зависимости

На рис. 17.6 приведены фрагменты выводимых данных, сгенерированные программой после выполнения шага 5б.

Model Summary

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .822[a] | .675 | .655 | 1.59269 |

a. Predictors: (Constant), ТРЕВ2, ТРЕВ

ANOVA[b]

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 173.929 | 2 | 86.964 | 34.283 | .000[a] |
| | Residual | 83.710 | 33 | 2.537 | | |
| | Total | 257.639 | 35 | | | |

a. Predictors: (Constant), ТРЕВ2, ТРЕВ

b. Dependent Variable: ТЕСТ

Coefficients[a]

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | .161 | 1.ББ9 | | .097 | .924 |
| | ТРЕВ | 4.490 | .629 | 3.633 | 7.139 | .000 |
| | ТРЕВ2 | -.338 | .055 | -3.148 | -6.185 | .000 |

a. Dependent Variable: ТЕСТ

Рис. 17.6. Фрагмент окна вывода после выполнения шага 5б

# Терминология, используемая при выводе

Термины, используемые в приведенных результатах, имеют тот же смысл, что и в предыдущем случае. Обратите внимание на значительный коэффициент корреляции зависимой переменной тест с независимыми переменными трев и трев2. Величина $R^2 = 0{,}675$ означает, что 67,5 % дисперсии переменной тест обусловлено влиянием со стороны переменных трев и трев2. Значения $F$-критерия и соответствующие значимости (для $F$ и $t$) говорят о сильном воздействии на зависимую переменную как обеих независимых переменных, так и каждой переменной в отдельности. Вопросы влияния нескольких независимых переменных будут рассмотрены в следующей главе. Как вы можете видеть, значения p (Beta) для нелинейных отношений не ограничиваются диапазоном от -1 до 1. В столбце B последней таблицы приведены коэффициенты уравнения регрессии. Сравните эти коэффициенты с коэффициентами уравнения для квадратичной регрессии, приведенного в середине главы.

# 18 Множественный регрессионный анализ



Множественная регрессия является расширением простой линейной регрессии, описанной в главе 17. С помощью простой регрессии оценивалась степень влияния одной независимой переменной (предиктора) на зависимую переменную (критерий). В отличие от простой регрессии множественная регрессия исследует влияние двух и более предикторов на критерий.

Анализ регрессии можно свести к геометрической интерпретации. Когда вычислена простая корреляция между двумя переменными, можно построить *линию регрессии* (линию «наилучшего соответствия»). Эта линия строится на основе уравнения регрессии; ее угловой коэффициент равен коэффициенту при независимой переменной, а сдвиг вверх по вертикальной оси определяется константой. Далее мы продемонстрируем множественную регрессию как последовательное усложнение простого регрессионного уравнения. Специально для этой главы нами создан файл данных help.sav. Этот файл содержит данные психологического исследования склонности людей оказывать помощь своим знакомым. Хотя данные являются вымышленными, результаты их обработки близки к результатам одного из реальных исследований.

## Уравнение множественной регрессии

Итак, в этой главе мы используем новый набор данных help.sav. Переменная помощь представляет время (в секундах), потраченное человеком на оказание помощи своему партнеру, и ее значения имеют нормальное распределение (среднее

равно 30, стандартное отклонение — 10). Переменная симпатия отражает оценку симпатии к партнеру в баллах от 1 до 20. На примере этих двух переменных мы продемонстрируем простую регрессию. В качестве зависимой выступит переменная помощь, а в качестве независимой — переменная симпатия (предполагается, что симпатия и сочувствие заставляют человека оказывать помощь, а не наоборот). Как показал анализ, коэффициент корреляции между переменными помощь и симпатия составляет 0,416 при значимости $p = 0{,}004$, что говорит о значительной связи между этими переменными. Константа и коэффициент регрессии составили, соответственно, 14,739 и 1,547. Таким образом, уравнение регрессии имеет следующий вид:

$$\text{помощь}_{\text{прогноз}} = 14{,}739 + 1{,}547\ (\text{симпатия}).$$

Если для некоторого испытуемого значение переменной симпатия составит 16, то на основе регрессионного уравнения мы можем прогнозировать, что переменная помощь примет следующее значение:

$$14{,}739 + 1{,}547\ (16) = 39{,}5.$$

Значение 16 сравнительно велико в соответствии с принятой шкалой; то же самое верно и в отношении прогнозируемого значения помощь, которое превышает среднее значение этой переменной почти на одно стандартное отклонение.

Аналогичные расчеты можно выполнять и при множественном регрессионном анализе: Различие заключается лишь в том, что при множественном анализе уравнение регрессии включает более чем одну зависимую переменную.

Помимо переменной симпатия с переменной помощь коррелируют и другие переменные файла help.sav. В частности, это переменные агрессия (агрессивность человека по отношению к партнеру, измеренная в баллах от 1 до 20) и польза (самооценка собственной полезности в баллах от 1 до 20). Множественный регрессионный анализ показал следующие коэффициенты при каждой из переменных: $B_{(\text{симпатия})} = 1{,}0328$, $B_{(\text{агрессия})} = 1{,}1676$, $B_{(\text{польза})} = 1{,}2569$, *константа* = -5,3147. Уравнение регрессии для множественного анализа имеет следующий вид:

$$\text{помощь}_{\text{прогноз}} = -5{,}3147 + 1{,}0328\ (\text{симпатия}) + 1{,}1676\ (\text{агрессия}) + 1{,}2569\ (\text{польза}).$$

Возьмем объект с номером 7 и рассчитаем для него прогнозируемое значение неременной помощь:

$$\text{помощь}_{\text{прогноз}} = -5{,}3147 + 1{,}0328\ (2) + 1{,}1676\ (10) + 1{,}2569\ (9) - 19{,}74.$$

Таким образом, человек, имеющий низкий показатель симпатии и средние показатели агрессивности и самооценки полезности, должен, согласно прогнозу, оказывать незначительную помощь. Фактическое значение переменной помощь для объекта 7 составило 21, что свидетельствует о высокой точности нашего прогноза.

# Коэффициенты регрессии

Положительный коэффициент при независимой переменной говорит о том, что с возрастанием последней значение зависимой переменной также возрастает. Верно и противоположное утверждение: при отрицательном коэффициенте с возрастанием значения независимой переменной значение зависимой переменной убывает. Тем не менее в большинстве исследований соотношение коэффициентов не позволяет делать вывод о воздействии того или иного фактора на зависимую переменную, поскольку независимые переменные, как правило, измеряются в разных шкалах и имеют разный масштаб. Чтобы разрешить эту проблему, был введен коэффициент регрессии р, принимающий значения от -1 до 1. Он отражает *частную корреляцию* независимой и зависимой переменных. Под частной корреляцией понимается воздействие, оказываемое на зависимую переменную со стороны одной независимой переменной при фиксированных значениях других независимых переменных (с учетом влияния последних). Чем больше данная независимая переменная коррелирует с другими независимыми переменными, тем меньше абсолютная величина ее коэффициента р. Так, для нашего примера с тремя независимыми переменными коэффициент р для переменной агрессия — это ее корреляция с зависимой переменной *после* исключения влияния переменных симпатия и польза. Для простой регрессии с одной независимой переменной коэффициент р равен коэффициенту парной корреляции зависимой и независимой переменных. Таким образом, коэффициент р является универсальной мерой влияния независимой переменной; его часто называют *стандартным коэффициентом регрессии*. И именно стандартные коэффициенты регрессии позволяют соотносить независимые переменные по их значению для оценки зависимой переменной.

Обратимся к записанному выше уравнению регрессии. Как мы можем видеть, все три независимые переменные входят в него с положительными коэффициентами. Такой результат вполне логичен для переменных симпатия и польза, однако вызывает недоумение то, что рост агрессивности субъекта влечет за собой увеличение его стремления оказывать помощь. Причина появления такого странного результата — хороший повод для дискуссии. Каждый раз, когда вы будете попадать в подобные «непредвиденные» ситуации, проверяйте правильность кодирования и ввода данных. Помните, что программа способна лишь генерировать результаты анализа, но в отношении их интерпретации она практически беспомощна.

# Коэффициент детерминации и пошаговые методы

Коэффициент *R* является мерой связи всей совокупности независимых переменных и зависимой переменной. Часто его называют *коэффициентом множественной корреляции*. Величина $R^2$ равна доле дисперсии зависимой переменной, обусловленной влиянием со стороны независимых переменных, и называется *коэффициентом детерминации*. Для регрессионного анализа с тремя независимыми переменными,

о котором шла речь, значение $R$ составило 0,571, а $R^2$ — 0,326. Это означает, что 32,6 % дисперсии переменной помощь определяется совокупным воздействием переменных агрессия, симпатия и польза. Дополнительная информация о коэффициентах множественной корреляции и детерминации приведена в разделе «Представление результатов».

Множественный регрессионный анализ позволяет использовать любое количество предикторов, но присутствие большого числа независимых переменных не всегда удобно. Было бы предпочтительно иметь в качестве предикторов как можно больше переменных, оказывающих значимое влияние на критерий, и как можно меньше переменных, не оказывающих такого влияния. В процедуру множественной регрессии SPSS включены методы, позволяющие производить пошаговый отбор в регрессионное уравнение только значимых независимых переменных. Одним из них является метод Forward (Прямой), суть которого заключается в следующем. Сначала процедура вычисляет, какая из независимых переменных имеет наибольший коэффициент корреляции с зависимой переменной, а затем составляет уравнение регрессии с участием этой переменной. Далее из числа оставшихся предикторов выбирается тот, который имеет наибольший коэффициент $\beta$, при условии, что $\beta$ является значимым. Выбранный предиктор включается в уравнение регрессии. Процесс продолжается до тех пор, пока не будут выбраны все предикторы, оказывающие значимое воздействие на зависимую переменную (имеющие статистически достоверные коэффициенты P). По умолчанию SPSS продолжает выбирать независимые переменные до тех пор, пока уровень значимости ($p$) коэффициентов р не превысит значения 0,05. Разумеется, при желании вы можете изменить величину порогового уровня значимости.

# Условия получения приемлемых результатов анализа

Регрессия, как и корреляция, анализирует линейные зависимости. В предыдущей главе мы рассмотрели процедуру оценки криволинейных зависимостей в контексте простого регрессионного анализа. Большая часть того, что было справедливо для простой регрессии, оказывается применимым и для множественного анализа. Если теория или статистический расчет показывает, что между критерием и одним или несколькими предикторами существует криволинейная зависимость, то можно ввести в качестве дополнительного предиктора, к примеру, квадрат какой-либо из независимых переменных. О том, что для этого нужно сделать, подробно рассказано в предыдущей главе.

В заключение обзора множественной регрессии рассмотрим основные условия, выполнение которых способствует получению действительно ценных и концептуально осмысленных результатов анализа.

- Ваше исследование должно быть продумано по форме и исполнению. Анализ регрессии для не связанных по смыслу величин приводит к бесполезным

результатам. К сожалению, описание тех условий, которые делают исследование «продуманным но форме и исполнению», выходит за рамки темы данной книги.

- Для того чтобы существующие корреляции были признаны значимыми, необходимо иметь достаточные размеры выборок. Трудно указать точные границы «достаточно большой выборки», однако, как правило, проблемы со значимостью начинают возникать при $N < 50$. Разумеется, чем большее число переменных вы привлекаете для анализа, тем больший размер выборки требуется для получения значимых результатов.
- Ваши данные должны быть корректными и не содержать ошибок.
- Распределение значений предикторов должно быть близким к нормальному. Желательно, чтобы значения асимметрий и эксцессов по модулю не превосходили 1. Тем не менее можно получить весьма точные результаты, если это требование не выполняется строго для каждого из предикторов, и даже в случае если в анализ входит дискретная переменная с небольшим числом значений. Нормальность распределения зависимой переменной также желательна, однако допустимы как отклонения от нормальности, так и использование дискретных переменных с малым числом значений.
- Наиболее жестким требованием является запрет па использование зависимых переменных, корреляции между которыми близки к 1 (-1). Не следует задействовать предикторы, схожие между собой но смыслу. Перед проведением регрессионного анализа никогда не бывает лишним вычисление корреляций между предикторами. В случае если коэффициент корреляции превысит значение 0,5, не рекомендуется одновременное присутствие соответствующих переменных в регрессионном анализе в качестве предикторов. В противном случае результаты исследования будут иметь сомнительную ценность.

Множественный регрессионный анализ является весьма сложной процедурой. И нескольких страниц краткого обзора явно недостаточно для его изучения и исчерпывающего понимания. Поэтому желательно, чтобы перед проведением множественного регрессионного анализа у вас уже были более глубокие знания о нем.

После беглого обзора простого и множественного регрессионного анализа мы можем приступить к его практической реализации с помощью программы SPSS, а также к интерпретации результатов.

# Пошаговые алгоритмы вычислений

Файл данных, который мы будем использовать в приводимых примерах, называется help.sav. Число объектов в этом файле равно 46, то есть $N = 81$. Далее перечислены те переменные файла, которые мы будем использовать.

- помощь — зависимая переменная, интерпретируемая как время (в секундах) оказания помощи партнеру (среднее — 30, стандартное отклонение — 10);

- симпатия — оценка своей симпатии к партнеру, нуждающемуся в помощи (по 20-балльной шкале);
- агрессия — оценка своей агрессивности к партнеру (по 20-балльной шкале);
- польза — оценка пользы от своей помощи (по 20-баллыюй шкале);
- проблема — оценка серьезности проблемы своего партнера (по 20-балльной шкале);
- эмпатия — оценка эмпатии (склонности к сопереживанию) как результат тестирования (по 10-балльной шкале).

Для проведения множественного регрессионного анализа сначала необходимо выполнить три подготовительных шага (обратите внимание, что в данном случае мы работаем с файлом help.sav).

**Шаг 1** Создайте новый файл данных или подготовьте существующий. Все необходимые действия описаны в главе 3.

**Шаг 2** Запустите программу SPSS при помощи значка па рабочем столе или команды Пуск ▸ Программы ▸ SPSS for Windows ▸ SPSS 11.5 for Windows (Start ▸ Programs ▸ SPSS for Windows ▸ SPSS 11.5 for Windows) главного меню Windows. В открывшемся после запуска программы диалоговом окне SPSS for Windows щелкните па кнопке Cancel (Отмена).

После выполнения этого шага на экране появится окно редактора данных SPSS.

**Шаг 3** Откройте файл данных, с которым вы намерены работать (в нашем случае это файл help.sav). Если он расположен в текущей папке, то выполните следующие действия:

1. Выберите в меню File (Файл) команду Open ▸ Data (Открытие ▸ Данные) или щелкните па кнопке Open File (Открытие файла) панели инструментов.
2. В открывшемся диалоговом окне дважды щелкните на имени help.sav или введите его с клавиатуры и щелкните па кнопке OK.

Независимо от того, открыта программа SPSS только что или какие-то процедуры уже выполнялись, в верхней части главного окна должна присутствовать строка меню (она показана ниже). Пока строка меню присутствует на экране, доступны все команды анализа данных. При этом окно с данными видеть не обязательно.

File Edit View Data Transform Analyze Graphs Utilities Window Help

При работе со сводными таблицами или при редактировании диаграмм некоторые пункты строки меню могут исчезать или меняться. Чтобы вернуться к главному окну и стандартной строке меню, щелкните на кнопке свертывания или восстановления текущего окна.

После завершения шага 3 на экране должно присутствовать окно редактора данных.

**Шаг 4** В меню Analyze (Анализ) выберите команду Regression ▸ Linear (Регрессия ▸ Линейная регрессия). На экране появится диалоговое окно Linear Regression (Линейная регрессия), представленное на рис. 18.1.

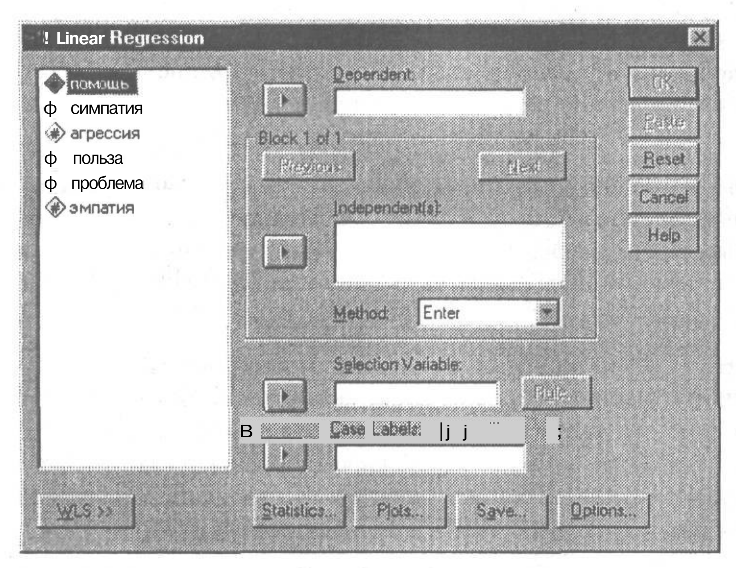



**Рис. 18.1.** Диалоговое окно Linear Regression

В верхней части диалогового окна расположено поле Dependent (Зависимая переменная), предназначенное для указания единственной переменной-критерия. Ниже следует список Independent(s) (Независимые переменные), в который будут помещены одна или несколько независимых переменных. Число последних теоретически не ограничено. С помощью кнопки Next (Следующий), расположенной справа от метки Block 1 of 1 (Блок 1 из 1), вы можете задать не один, а несколько наборов предикторов и, тем самым, провести одновременно несколько вариантов регрессионного анализа. Как только вы определите все независимые переменные и установите необходимые параметры анализа, щелчок на кнопке Next (Следующий) приведет к очистке списка Independent(s) (Независимые переменные), который будет готов к принятию нового набора предикторов. После того как вы зададите все блоки независимых переменных и щелкнете на кнопке (Ж, SPSS выполнит все варианты регрессионного анализа и сгенерирует результаты для каждого из них. Обратите внимание, что во всех вариантах анализа, проводимых одновременно, должна быть общая зависимая переменная. В противном случае процедуру регрессии каждый раз пришлось бы запускать заново.

Важным элементом диалогового окна Linear Regression (Линейная регрессия) является раскрывающийся список Method (Метод). Пункты этого списка определяют алгоритмы включения независимых переменных в уравнение регрессии.

- Enter (Включение) — метод, применяемый по умолчанию. Все независимые переменные включаются в уравнение независимо от степени их корреляции с переменной-критерием.
- Forward (Прямой) —пошаговое включение переменных с проверкой на значимость их частной корреляции с критерием. В результате в уравнение включаются все переменные, имеющие значимую частную корреляцию с переменной-критерием. Включение производится в порядке возрастания *p*-уровня.
- Backward (Обратный) — пошаговый метод, сначала включающий в уравнение регрессии все независимые переменные, а затем поочередно удаляющий все переменные, чья корреляция с критерием имеет уровень значимости выше заданного порогового значения. Как правило, пороговым значением является $p = 0{,}1$.
- Stepwise (По шагам) — комбинация пошаговых методов Forward (Прямой) и Backward (Обратный). Основной идеей является изменение доли влияния независимой переменной на критерий при появлении в уравнении других независимых переменных. Если влияние какой-либо из включенных переменных становится слишком слабым, она исключается из уравнения. Подобный метод используется при регрессионном анализе наиболее часто.
- Remove (Исключение) — это метод принудительного удаления переменных. Он требует предварительного задания метода Enter (Включение) в качестве предыдущего блока, например, Block 1 of 1 (Блок 1 из 1). При задании следующего блока, в данном случае Block 2 of 2 (Блок 2 из 2), в список Independent(s) (Независимые переменные) вы сможете ввести те независимые переменные, которые хотите исключить из уравнения регрессии. При выполнении команды вы получите результат со всеми заданными переменными, а затем — результат с удаленными переменными. Если в анализе участвуют несколько блоков, то можно задавать операцию удаления после каждого из них.

Кнопка WLS>> позволяет выбрать переменную (из числа тех, которые не входят в анализ в качестве предикторов) для использования в роли весового коэффициента какого-либо из предикторов, но требуется это очень редко.

Кнопка Plots (Диаграммы) используется для графического отображения остатков (см. главу 26).

Кнопка Statistics (Статистики) открывает диалоговое окно Linear Regression: Statistics (Линейная регрессия: Статистики), представленное на рис. 18.2. По умолчанию в окне установлены два флажка. Флажок Estimates (Оценки) включает в вывод коэффициенты *B*, стандартные коэффициенты регрессии $\beta$, а также соот-

ветствующие стандартные ошибки, $t$-критерии и уровни значимости. Флажок Model fit (Характеристики модели) генерирует значения множественного коэффициента корреляции $R$, величину $R^2$, таблицу дисперсионного анализа (ANOVA), соответствующие $F$-величины и уровни их значимости. Таким образом, указанные флажки отвечают за генерацию основных элементов регрессионного анализа.

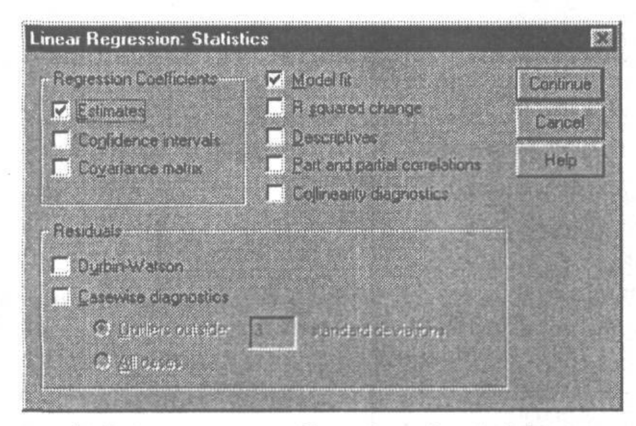



**Рис. 18.2.** Диалоговое окно Linear Regression: Statistics

Ниже перечислены наиболее важные флажки диалогового окна Linear Regression: Statistics (Линейная регрессия: Статистики).

- Confidence intervals (Доверительные интервалы) — включает в вывод для коэффициентов *B* доверительный интервал в 95 %.
- Covariance matrix (Ковариационная матрица) — генерирует таблицу, под главной диагональю которой расположены ковариации, на главной диагонали — дисперсии, а над главной диагональю — корреляции.
- R squared change (Изменение квадрата R) — для методов Forward (Прямой) и Stepwise (По шагам) указывает изменения коэффициента $R^2$ при введении новых переменных в уравнение регрессии.
- Descriptives (Описательные статистики) — включает средние значения переменных, стандартные отклонения, а также корреляционную матрицу.
- Collinearity diagnostics (Анализ коллинеарности) — устанавливает наличие коллинеарности (корреляций, близких к 1) между переменными.

Щелчок на кнопке Save (Сохранение) в диалоговом окне Linear Regression (Линейная регрессия) приводит к открытию диалогового окна Linear Regression: Save (Линейная регрессия: Сохранение), показанного на рис. 18.3. В нем имеется множество флажков, названия которых, как правило, представляют загадку даже для людей, искушенных в математике. Однако некоторые весьма полезны.

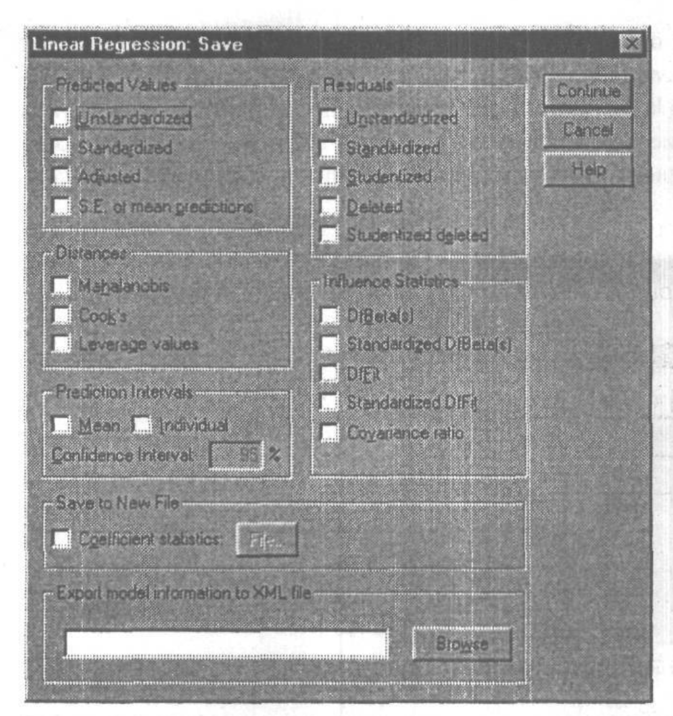



**Рис. 18.3.** Диалоговое окно Linear Regression: Save

Данное окно позволяет создать в файле данных новые переменные, содержащие значения, соответствующие установленным флажкам.

- В группе Predicted Values (Прогнозируемые значения) имеется 4 флажка. Флажок Unstandardized (Нестандартизированные значения) генерирует прогнозируемые значения, которые бывает полезно сравнить с фактическими значениями для оценки адекватности уравнения регрессии. Кроме того, этот флажок позволяет получать прогнозы (оценки) зависимой переменной для тех объектов, для которых ее истинные значения неизвестны. Флажок Standardized (Стандартизованные значения) позволяет рассчитывать прогнозируемые значения для стандартизованных распределений выбранных переменных.
- В группу Residuals (Остатки) включены 5 флажков, речь о которых пойдет в главе 26.
- Флажки в группе Influence Statistics (Показатели влияния) позволяют исключать из выборки те или иные объекты. Так, если в команде спортсменов-бегунов один пробегает дистанцию гораздо хуже или гораздо лучше других, его результаты значительно искажают статистические показатели всей команды. Иногда подобные значения («выбросы») желательно исключать из анализа. К сожалению, подробное изложение этой процедуры выходит за пределы темы данной книги.

- *С* помощью поля и флажков из группы Prediction Intervals (Интервалы прогнозируемости) можно изменять число процентов в доверительном интервале для средних или отдельных значений (по умолчанию — 95 %).
- Флажки в группе Distances (Расстояния) предоставляют три способа измерения расстояния между объектами (см. главу 22).

Наконец, рассмотрим диалоговое окно Linear Regression: Options (Линейная регрессия: Параметры), представленное на рис. 18.4. Это окно открывается при щелчке на кнопке Options (Параметры) в окне Linear Regression (Линейная регрессия).

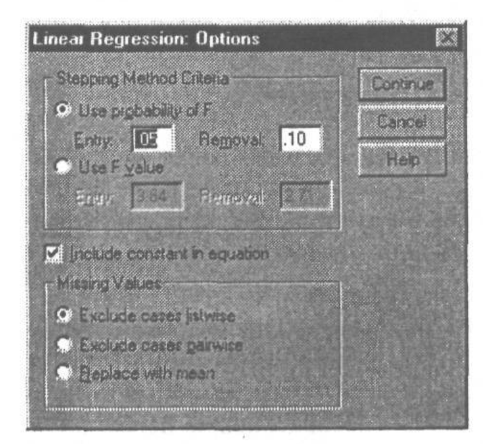



**Рис. 18.4.** Диалоговое окно Linear Regression: Options

Флажок Include constant in equation (Включать константу в уравнение) установлен по умолчанию, и без веских на то причин сбрасывать его не рекомендуется. Переключатели в группе Stepping Method Criteria (Критерии пошаговых методов) позволяют управлять выполнением методов Forward (Прямой), Backward (Обратный) и Stepwise (По шагам). В поле Entry (Включение) вы можете указать пороговую величину значимости для включения переменных в уравнение регрессии методами Forward (Прямой) и Stepwise (По шагам), а в поле Removal (Исключение) — для исключения переменных методами Backward (Обратный) и Stepwise (По шагам); в первом по умолчанию уставлено значение 0,05, во втором — значение 0,1. Если в качестве критерия отбора желательно использовать не значимость *F*-критерия, а саму его величину, следует установить переключатель Use F Value (Использовать значение *F*-критерия) и при необходимости указать пороговые значения в полях Entry (Включение) и Removal (Удаление). Группа переключателей Missing Values (Пропущенные значения) позволяет выбрать способ обработки отсутствующих значений (см. раздел «Обработка пропущенных значений» в главе 4).

Далее приведены два примера (шаги 5 и 5а), в первом из которых проводится множественный регрессионный анализ с участием зависимой переменной помощь и пяти предикторов симпатия, проблема, эмпатия, польза и агрессия. Для составления уравнения регрессии мы воспользуемся методом Forward (Прямой). Во втором примере мы используем некоторые из описанных параметров.

Шаг 5 После выполнения шага 4 у вас должно быть открыто диалоговое окно Linear Regression (Линейная регрессия), показанное на рис. 18.1.

1. Щелкните сначала на переменной помощь, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в поле Dependent (Зависимая переменная).
2. Щелкните сначала на переменной симпатия, чтобы выделить ее, а затем — на второй сверху кнопке со стрелкой, чтобы переместить переменную в список Independent(s) (Независимые переменные).
3. Повторите предыдущее действие для переменных проблема, эмпатия, польза и агрессия.
4. В раскрывающемся списке Method (Метод) выберите пункт Forward (Прямой).
5. Щелкните на кнопке OK, чтобы открыть окно вывода.

В результате программа сгенерирует данные, показывающие, какая из независимых переменных оказывает наибольшее влияние на зависимую переменную. Метод Forward (Прямой) обеспечит включение в уравнение регрессии всех предикторов, имеющих значимую частную корреляцию с критерием р в порядке убывания значимости.

В следующем примере мы проведем регрессионный анализ с участием тех же переменных, что и в предыдущем, однако будем использовать метод Stepwise (По шагам), включим в результат статистики для коэффициентов *B*, описательные статистики и характеристики модели, а также добавим новую переменную и зададим пороговые значения значимости включения и исключения предикторов.

Шаг 5а После выполнения шага 4 должно быть открыто диалоговое окно Linear Regression (Линейная регрессия), показанное на рис. 18.1. Если вы уже успели поработать с этим окном, очистите его щелчком на кнопке Reset (Сброс) и выполните следующие действия.

1. Щелкните сначала на переменной помощь, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в поле Dependent (Зависимая переменная).
2. Щелкните сначала на переменной симпатия, чтобы выделить ее, а затем — на второй сверху кнопке со стрелкой, чтобы переместить переменную в список Independent(s) (Независимые переменные).
3. Повторите предыдущее действие для переменных проблема, эмпатия, польза и агрессия.
4. В раскрывающемся списке Method (Метод) выберите пункт Stepwise (По шагам).
5. Щелкните на кнопке Statistics (Статистики), чтобы открыть диалоговое окно Linear Regression: Statistics (Линейная регрессия: Статистики), показанное на рис. 18.2.
6. Установите флажок Descriptives (Описательные статистики) и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Linear Regression (Линейная регрессия).

7. Щелкните на кнопке Save (Сохранение), чтобы открыть диалоговое окно Linear Regression: Save (Линейная регрессия: Сохранение), показанное на рис. 18.3.
8. Установите флажок Unstandardized (Нестандартизированные значения) и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Linear Regression (Линейная регрессия).
9. Щелкните на кнопке Options (Параметры), чтобы открыть диалоговое окно Linear Regression: Options (Линейная регрессия: Параметры), показанное на рис. 18.4.
10. В поле Entry (Включение) введите значение 0,1, нажмите клавишу Tab, в поле Removal (Удаление) введите значение 0,2 и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Linear Regression (Линейная регрессия).
11. Щелкните на кнопке OK, чтобы открыть окно вывода.

В результате выполнения приведенных инструкций будут сгенерированы данные, позволяющие судить о том, какая из независимых переменных оказывает наибольшее влияние на критерий. При составлении уравнения регрессии сначала в него включаются переменные, чья частная корреляция (β) с зависимой переменной имеет уровень значимости не выше 0,1. Если затем обнаружится, что из включенных переменных какие-либо обнаруживают новый уровень значимости, превышающий значение 0,2, то они исключаются из уравнения. Кроме того, в результате выполнения процедуры будет создана переменная для хранения прогнозируемых значений переменной помощь, рассчитанных по составленному уравнению регрессии. В выводе вы также сможете найти корреляционную матрицу для всех переменных и описательные статистики.

После выполнения шага 5 программа автоматически активизирует окно вывода. Для просмотра результатов вы при необходимости можете воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

# Печать результатов и выход из программы

Далее описана типичная процедура печати результатов статистического анализа (или нескольких анализов). После выполнения шага 5 должно быть открыто окно вывода.

Шаг 6 В окне вывода укажите фрагменты, выводимые на печать (см. раздел «Окно вывода» в главе 2), в меню File (Файл) выберите команду Print (Печать), при необходимости задайте параметры печати и щелкните на кнопке OK.

Последнее, что необходимо сделать после завершения исследования и печати результатов, — это выйти из программы SPSS.

**Шаг 7** Для выхода из программы в меню File (Файл) выберите команду Exit (Выход).

Иногда после выполнения команды Exit (Выход) на экране могут появляться небольшие диалоговые окна с вопросом о необходимости сохранения сделанных в файлах изменений и кнопками, описывающими возможные варианты ответа. Для завершения работы просто щелкайте на соответствующих кнопках.

# Представление результатов

## Множественный регрессионный анализ

Представленные на рис. 18.5 данные генерируются программой при выполнении шага 5. В таблицах представлены не все, а только окончательные результаты. Визуальная форма некоторых таблиц для удобства их восприятия немного изменена без ущерба для содержания.

**Variables Entered/Removed**

| | Model |
|---|---|
| | 3 |
| Method | Forward (Criterion: Probability-of-F-to-enter <= .050) |

a. Dependent Variable: ПОМОЩЬ

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 3 | .571[c] | .326 | .278 | 8.622 |

c. Predictors: (Constant), СИМПАТИЯ, ПОЛЬЗА, АГРЕССИЯ

**ANOVA**

| | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Regression | 1512.813 | 3 | 504.271 | 6.784 | .001 |
| Residual | 3121.991 | 42 | 74.333 | | |
| Total | 4634.804 | 45 | | | |

**Coefficients**

| | Unstandardized Coefficients | | Standardized Coefficients | | |
|---|---|---|---|---|---|
| | B | Std. Error | Beta | t | Sig. |
| (Constant) | -5.315 | 8.075 | | -.658 | .514 |
| СИМПАТИЯ | 1.033 | .500 | .278 | 2.065 | .045 |
| ПОЛЬЗА | 1.257 | .603 | .276 | 2.083 | .043 |
| АГРЕССИЯ | 1.168 | .567 | .269 | 2.059 | .046 |

**Excluded Variables**

| | Beta In | t | Sig. | Partial Correlation | Collinearity Statistics Tolerance |
|---|---|---|---|---|---|
| ПРОБЛЕМА | .162 | 1.136 | .262 | .175 | .779 |
| ЭМПАТИЯ | .116 | .814 | .420 | .126 | .792 |

**Рис. 18.5.** Фрагменты окна вывода после выполнения шага 5

Как можно видеть, из пяти предикторов в уравнение регрессии входят лишь три: симпатия, агрессия и польза. Коэффициент множественной корреляции *R* отражает связь зависимой переменной помощь с совокупностью независимых переменных и равен 0,571. Значение $R^2$ составляет 0,326 и показывает, что 32,6 % дисперсии переменной помощь обусловлено влиянием предикторов. Стандартные коэффициенты регрессии р отражают относительную степень влияния каждого из предикторов; для переменной симпатия р - 0,278, а для переменных агрессия и польза соответственно р - 0,276 и р = 0,269. Каждая из независимых переменных имеет положительную корреляцию с зависимой переменной.

## Изменение показателей при добавлении новых переменных

Представленные на рис. 18.6 результаты генерируются программой при выполнении шага 5а. Как и в предыдущем случае, мы максимально адаптировали их для лучшего восприятия.

**Model Summary[d]**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .416[a] | .173 | .154 | 9.335 |
| 2 | .508[b] | .258 | .224 | 8.941 |
| 3 | .571[c] | .326 | .278 | 8.622 |

a. Predictors: (Constant), СИМПАТИЯ

b. Predictors: (Constant), СИМПАТИЯ, ПОЛЬЗА

c. Predictors: (Constant), СИМПАТИЯ, ПОЛЬЗА, АГРЕССИЯ

**ANOVA**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 800.289 | 1 | 800.289 | 9.183 | .004 |
| | Residual | 3834.515 | 44 | 87.148 | | |
| | Total | 4634.804 | 45 | | | |
| 2 | Regression | 1197.635 | 2 | 598.817 | 7.491 | .002 |
| | Residual | 3437.170 | 43 | 79.934 | | |
| | Total | 4634.804 | 45 | | | |
| 3 | Regression | 1512.813 | 3 | 504.271 | 6.784 | .001 |
| | Residual | 3121.991 | 42 | 74.333 | | |
| | Total | 4634.804 | 45 | | | |

Рис. 18.6. Фрагменты окна вывода после выполнения шага 5а

Полученные данные показывают, каким образом изменялись характеристики регрессионного анализа при поэтапном включении новых предикторов в уравнение регрессии.

С возрастанием числа предикторов значения *R*, $R^2$ и исправленная величина $R^2$ возрастают, в то время как значение стандартной ошибки убывает. Та же тенденция наблюдается и в отношении числа степеней свободы, суммы квадратов и среднего квадратов.

# Терминология, используемая при выводе

Ниже дана трактовка терминов, используемых программой в окне вывода.

- Probability-of-F-to-enter (Значимость $F$-включения) — максимальный уровень значимости переменных, вводимых в уравнение регрессии, в данном случае равный $p$ - 0,050.
- R — коэффициент множественной корреляции, отражающий связь совокупности предикторов симпатия, агрессия и польза с критерием помощь.
- R Square (Квадрат R) — коэффициент детерминации ($R^2$), равный доле дисперсии зависимой переменной помощь, обусловленной влиянием независимых переменных симпатия, агрессия и польза.
- Adjusted R Square (Исправленный квадрат R) — исправленная величина $R^2$. Величина $R^2$, используемая в расчетах, на практике оказывается несколько завышенной. Исправленная величина $R^2$ ближе к реальным результатам.
- Std. Error (Стандартная ошибка) — в таблице Model Summary стандартное отклонение ожидаемого значения переменной помощь. Как видно из выводимых данных, с введением каждой повой независимой переменной в уравнение регрессии эта величина уменьшается.
- Regression (Регрессия) — статистика, отражающая влияние предикторов на зависимую переменную.
- Residual (Остаток) — статистика, отражающая внешнее (не обусловленное предикторами) влияние на независимую переменную.
- df (Число степеней свободы) — для регрессии это число равно числу независимых переменных в уравнении регрессии. Для остатка это число равно разности размера выборки и числа независимых переменных, входящих в уравнение регрессии, уменьшенной на единицу (46 - 3 - 1 = 42).
- Sum of Squares (Сумма квадратов) — для регрессии это число равно межгрупповой сумме квадратов в дисперсионном анализе. Для остатка это число равно внутригрупповой сумме квадратов в дисперсионном анализе. Обратите внимание, что в данном примере воздействие предикторов ограничивается лишь небольшой долей дисперсии критерия; это также отражено в величине $R^2$.
- Mean Square (Средний квадрат) — отношение суммы квадратов к числу степеней свободы.
- F ($F$-критерий) — отношение среднего квадрата регрессии к среднему квадрату остатка.
- Sig. (Значимость) — вероятность того, что результат является случайным.

- B — коэффициенты и константа уравнения регрессии, связывающего критерий и предикторы:

  $помощь_{прогноз}$ - -5,3147 +1,0328 (симпатия) +1,1676 (агрессия) +1,2569 (польза).

- Std. Error (Стандартная ошибка) — в таблице Coefficients является мерой стабильности коэффициентов *B* и равна стандартному отклонению их значений, рассчитанных для большого числа выборок.
- Beta (Бета) — стандартизованный коэффициент регрессии (β), представляющий собой коэффициенты *B* для независимых переменных, представленных в г-шкале. Для линейных взаимодействий р по абсолютному значению не превосходит 1; для криволинейных взаимодействий это условие не является обязательным.
- t — отношение коэффициента *B* к своей стандартной ошибке.
- Beta in (Бета в уравнении) — значения коэффициента р для переменных, не включенных в уравнение регрессии в предположении, что они в него включены.
- Partial Correlation (Частная корреляция) — коэффициенты частной корреляции для переменных, входящих в уравнение регрессии. Наличие нескольких коррелирующих переменных в уравнении регрессии взаимно снижает их частную корреляцию.
- Tolerance (Толерантноаь) — мера коллинеарности. Определяется как 1 - $Ri^2$, где *Ri* — коэффициент множественной корреляции для случая, когда *i*-й предиктор выступает в качестве зависимой переменной. Если значение толерантности близко к 0, то соответствующий предиктор практически является линейной комбинацией остальных предикторов. Толерантность, близкая к 1, напротив, свидетельствует о независимости данного предиктора от остальных. Включение в анализ предикторов, между которыми имеется линейная зависимость, увеличивает величины стандартных ошибок, тем самым снижая надежность результатов. Кроме того, появляются и другие скрытые проблемы, такие как погрешности вычислений.

# 19 Анализ надежности



Анализ надежности, рассматриваемый в этой главе, имеет отношение к надежности измерений в психологии и других социальных науках. Этот анализ еще известен как анализ заданий (пунктов теста) и применяется для отбора наиболее пригодных вопросов или заданий измерительной методики (вопросника, анкеты, теста). Обычно при разработке методики сначала составляют ее предварительный вариант, включающий избыточное количество заданий (пунктов), который апробируется на достаточно представительной выборке респондентов. Затем проводят анализ надежности, который позволяет при помощи многочисленных критериев исключить неподходящие задания.

Мы коснемся двух показателей надежности: *коэффициента Кронбаха* (Chronbach), обозначаемого буквой а (альфа), и *надежности половинного расщепления* (split-half reliability). Существуют и другие типы надежности, однако мы ограничимся лишь их упоминанием в разделе пошаговых процедур, а основное внимание сосредоточим па наиболее употребительном коэффициенте а. После описаний двух видов надежности мы приведем примеры, иллюстрирующие их применение.

## Коэффициент альфа

Коэффициент а является мерой внутренней согласованности, или однородности, измерительной шкалы. Как правило, а лежит в пределах от 0 до 1, однако может принимать и отрицательные значения. Последние свидетельствуют о том, что часть элементов, или пунктов, шкалы измеряет противоположные величины. Чем ближе коэффициент а к 1, тем выше внутренняя согласованность системы заданий. Коэффициент а можно рассматривать как показатель корреляции между измеренными признаками для выборок равного объема. Если, к примеру, составить

10 тестов с сотней случайных вопросов из банка данных объемом в 1000 вопросов, то коэффициент а будет характеризовать корреляцию между этими тестами. Формула для вычисления коэффициента а выглядит следующим образом:

$$\alpha = \frac{kr}{1+(k-1)r}.$$

Здесь $k$ — число пунктов шкалы, $r$ — средний коэффициент корреляции между нарами пунктов.

Из приведенной формулы следует, что коэффициент а возрастает как с увеличением числа пунктов шкалы, так и с усилением корреляции между пунктами.

Применение а Кронбаха основано па модели, предполагающей наличие большей дисперсии у более надежного теста: чем надежнее тест, тем большая чувствительность (различительная способность) пунктов теста.

Если в методике применяются задания дихотомического типа («да» — «нет», «правильно» — «неправильно»), то формула для вычисления коэффициента а Кронбаха идентична так называемой формуле Кьюдера—Ричардсона (Kuder & Richardson).

# Надежность половинного расщепления

Надежность половинного расщепления вычисляется в случае, когда число элементов достаточно велико, чтобы можно было разделить шкалу па две половины. Рассмотрим тест 16 личностных свойств Раймонда Каттелла (Raymond Cattell). Тест состоит из 187 вопросов, из которых от 10 до 14 вопросов определяют одно из 16 свойств. Разумеется, было бы бессмысленно делить на две половины все 187 вопросов, поскольку в каждой половине оказалось бы множество вопросов, относящихся к разным свойствам, а следовательно, несопоставимых друг с другом. Более разумным подходом было бы разделить пополам совокупность вопросов, измеряющих одно свойство: например из 14 вопросов, определяющих тревожность, в случайном порядке выбрать 7 и сформировать первую группу, а из оставшихся — вторую группу. В качестве показателя надежности используется коэффициент корреляции Спирмена—Брауна (Spearman—Brown) между двумя образованными половинами.

# Пошаговые алгоритмы вычислений

Для проведения анализа надежности мы будем использовать файл данных TestA.sav, содержащий 12 переменных для 120 объектов (респондентов). Переменные имеют имена a1, ..., a12 и соответствуют пунктам разрабатываемой семантической

шкалы для оценки степени альтруизма. Каждая переменная — это оценка респондентом другого человека по наличию (1 — «да») или отсутствию (0 — «пет») у него одной из 12 характеристик. Список этих характеристик (прилагательных) вы можете найти в разделе «Представление результатов».

Чтобы приступить к анализу надежности, сначала нужно выполнить три подготовительных шага.

**Шаг 1** Создайте новый файл данных или подготовьте существующий. Все необходимые действия описаны в главе 3.

**Шаг 2** Запустите программу SPSS при помощи значка на рабочем столе или команды Пуск ▸ Программы ▸ SPSS for Windows ▸ SPSS 11.5 for Windows (Start ▸ Programs ▸ SPSS for Windows ▸ SPSS 11.5 for Windows) главного меню Windows. В открывшемся после запуска программы диалоговом окне SPSS for Windows щелкните на кнопке Cancel (Отмена).

После выполнения этого шага на экране появится окно редактора данных SPSS.

**Шаг 3** Откройте файл данных, с которым вы намерены работать (в нашем случае это файл TestA.sav). Если он расположен в текущей папке, то выполните следующие действия:

1. Выберите в меню File (Файл) команду Open ▸ Data (Открытие ▸ Данные) или щелкните на кнопке Open File (Открытие файла) панели инструментов.
2. В открывшемся диалоговом окне дважды щелкните на имени TestA.sav или введите его с клавиатуры и щелкните на кнопке OK.

Независимо от того, открыта программа SPSS только что или какие-то процедуры уже выполнялись, в верхней части главного окна должна присутствовать строка меню (она показана ниже). Пока строка меню присутствует на экране, доступны все команды анализа данных. При этом окно с данными видеть не обязательно.

File Edit View Data Transform Analyze Graphs Utilities Window Help

При работе со сводными таблицами или при редактировании диаграмм некоторые пункты строки меню могут исчезать или меняться. Чтобы вернуться к главному окну и стандартной строке меню, щелкните на кнопке свертывания или восстановления текущего окна.

После завершения шага 3 на экране должно присутствовать окно редактора данных.

**Шаг 4** В меню Analyze (Анализ) выберите команду Scale ▸ Reliability Analysis (Шкалирование ▸ Анализ надежности). Откроется диалоговое окно Reliability Analysis (Анализ надежности), показанное на рис. 19.1.

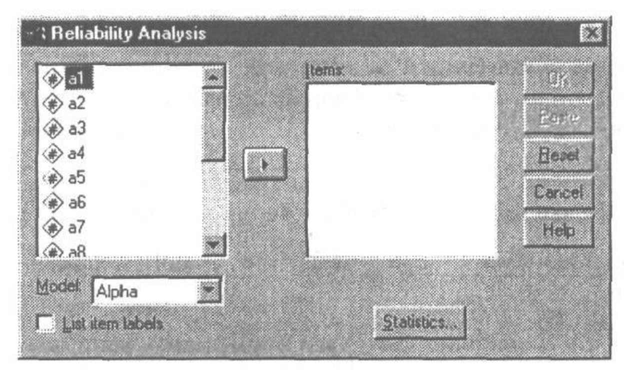



**Рис. 19.1.** Диалоговое окно Reliability Analysis

Параметры команды анализа надежности настраиваются при помощи двух диалоговых окон: основного, показанного на рис. 19.1, и диалогового окна Reliability Analysis: Statistics (Анализ надежности: Статистики), открывающегося при щелчке па кнопке Statistics (Статистика) и представленного па рис. 19.2,

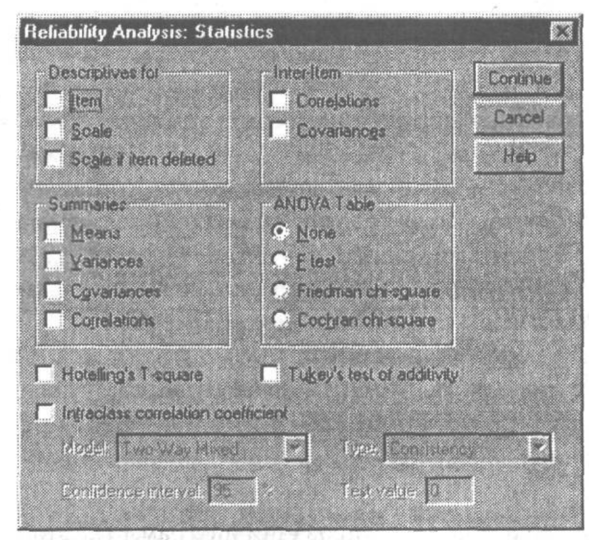



**Рис. 19.2.** Диалоговое окно Reliability Analysis: Statistics

Основное окно команды Reliability Analysis (Анализ надежности) почти не отличается от диалоговых окон большинства команд статистических процедур SPSS. Слева расположен список доступных переменных текущего файла данных, а справа находятся пять стандартных кнопок. Список Items (Пункты) предназначен для ввода имен переменных, включаемых в анализ надежности. Раскрывающийся список Model (Модель) позволяет задать модель анализа надежности, но умолчанию выбран пункт Alpha (Альфа). Далее приводятся описания остальных пунктов списка.

- Split-half (Половинное расщепление) — программа разбивает переменные на 2 группы в порядке их перечисления (число переменных в группах одинаково, если общее число переменных четно; в противном случае первая группа содержит на одну переменную больше) и сравнивает эти группы между собой.
- Guttman (Гутман) — анализ основан на операции вычисления нижних границ.
- Parallel (Параллель) — анализ проводится в предположении, что все переменные имеют равные дисперсии.
- Strict parallel (Строгая параллель) — анализ проводится в предположении, что все переменные имеют равные средние значения, дисперсии фактических значений и дисперсии ошибок.

Флажок List Item Labels (Список меток пунктов) позволяет включить в выводимые результаты список, в котором вместе с именами переменных будут перечислены их метки. Существует одна небольшая проблема: если вы используете имена переменных, оканчивающиеся числами (например, тест5, а9, тест12), и число этих переменных превышает 9, логический порядок их сортировки нарушается. Так, после имени тест1 вы можете увидеть имена тест10, тест11, тест12, и только потом тест2. Для того чтобы переменные располагались в нужном порядке (тест1, тест2 и т. д.), в меню Edit (Редактирование) выберите команду Options (Параметры). На экране появится большое диалоговое окно, содержащее множество вкладок и позволяющее настроить параметры программы под собственный вкус. В правой части вкладки General (Общие параметры) находится группа Variable Lists (Списки переменных) из двух пар переключателей. Нижняя пара переключателей определяет тип сортировки имен переменных: переключатель Alphabetical (Алфавитный) задает алфавитный порядок сортировки, а при установке переключателя File (Файл) имена переменных соответствуют исходному порядку следования в файле данных. Большинство исследователей предпочитают оставлять установленным переключатель File (Файл), поскольку при создании файла, как правило, порядок следования переменных логически продумывается.

Как уже упоминалось, при щелчке на кнопке Statistics (Статистики) открывается диалоговое окно Reliability Analysis: Statistics (Анализ надежности: Статистики), представленное на рис. 19.2. В нем вы можете задать параметры команды Reliability Analysis (Анализ надежности). Далее будут приведены описания всех флажков данного диалогового окна, за исключением тех, которые находятся в группе ANOVA Table (Таблица ANOVA). Дисперсионный анализ (ANOVA) и критерий $\chi^2$ рассматривались нами ранее в этой книге и, кроме того, эти виды анализа весьма редко используются в контексте анализа надежности. Остальные флажки перечислены в порядке их размещения в группах Descriptives for (Описательные статистики для), Summaries (Итоги), Inter-Item (Между пунктами). Кроме того, последни-

ми даны описания флажков Hotelling T-square (Т-квадрат Хотеллинга ) и Tukey's test of additivity (Критерий аддитивности Тьюки).

- Item (Пункт) — вычисление средних значений и стандартных отклонений переменных, участвующих в анализе.
- Scale (Шкала) — включение в вывод среднего значения, дисперсии, стандартного отклонения и размера выборки для суммы всех переменных шкалы. Под суммой переменных понимается сумма значений всех 12 переменных (пунктов) для каждого объекта. Таким образом обеспечивается вывод статистик для переменной, полученной в результате суммирования.
- Scale if item deleted (Шкала, если пункт удален) — включение в вывод для каждой переменной значения коэффициента а, подсчитанного для всей шкалы в предположении, что данная переменная исключена. В большинстве случаев этот флажок следует устанавливать.
- Means (Средние) — вычисление средних значений но всем объектам для каждой переменной. В вывод включаются все средние значения, минимальное и максимальное из этих значений, размах, отношение максимального среднего к минимальному, а также дисперсия средних значений. Несмотря на некоторую сложность, этот флажок позволяет получить важные для анализа надежности результаты.
- Variances (Дисперсии) — этот флажок аналогичен флажку Means (Средние) за тем исключением, что вычисления выполняются не для средних значений, а для дисперсий.
- Covariances (Ковариации) в группе Summaries (Итоги) — вычисление ковариаций между каждой переменной и суммой всех остальных переменных. В выводимые результаты также включаются средние значения полученных ковариаций, их минимум, максимум, размах, отношение максимума к минимуму и дисперсия.
- Correlations (Корреляции) в группе Summaries (Итоги) — этот флажок отличается от флажка Covariances (Ковариации) лишь тем, что вычисляются не ковариаций, а корреляции. Среднее корреляций для всех переменных является значением $r$ в формуле вычисления коэффициента а.
- Correlations (Корреляции) в группе Inter-Item (Между пунктами) — включение в вывод корреляционной матрицы для всех переменных, участвующих в анализе.
- Covariances (Ковариации) в группе Inter-Item (Между пунктами) — создание ковариационной матрицы для всех переменных, участвующих в анализе.
- Hotelling T-square (Т-квадрат Хотеллинга) — применение множественного сравнения по $t$-критерию для проверки достоверности различий средних значений всех переменных, участвующих в анализе.

- Tukey's test of additivity (Критерий аддитивности Тьюки) — применение критерия для проверки линейности зависимости между переменными, участвующими в анализе.

Далее приведены три примера, иллюстрирующие применение критерия надежности. На шаге 5 проводится анализ надежности с параметрами, заданными по умолчанию. На шаге 5а проводится анализ надежности с применением некоторых из описанных параметров. Данные, полученные в результате выполнения шага 5, показывают, что для 12 переменных надежность весьма невелика. Поэтому из списка на шаге 5а удаляются 3 переменные, нарушающие внутреннюю согласованность шкалы; тем самым повышается ее надежность. Наконец, на шаге 5б проводится анализ надежности половинного расщепления с 9 из 12 исходных переменных. Результаты всех трех вариантов анализа включены в раздел «Представление результатов».

В следующем примере проводится анализ надежности с набором параметров, заданных по умолчанию.

**Шаг 5** После выполнения шага 4 должно быть открыто диалоговое окно Reliability Analysis (Анализ надежности), показанное на рис. 19.1. Если вы уже успели поработать с этим окном, щелкните на кнопке Reset (Сброс).

1. Щелкните сначала на переменной a1, чтобы выделить ее, а затем — на кнопке со стрелкой, чтобы переместить переменную в список Items (Пункты).
2. Повторите предыдущее действие для переменных a2 ... a12.
3. Установите флажок List item labels (Список меток пунктов).
4. Щелкните на кнопке Statistics (Статистики), чтобы открыть диалоговое окно Reliability Analysis: Statistics (Анализ надежности: Статистики), показанное на рис. 19.2.
5. В группе Descriptives for (Описательные статистики для) установите флажок Scale if item deleted (Шкала, если элемент удален). Щелкните па кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Reliability Analysis (Анализ надежности).
6. Щелкните на кнопке OK, чтобы открыть окно вывода.

При выполнении этого шага все переменные попадают в список Items (Пункты) в том порядке, в котором они были в исходном файле. В вывод вместе с именами переменных включаются метки, представляющие соответствующие вопросы теста, а также коэффициент а и объем выборки *N*. Если вопросы теста вам видеть не требуется, флажок List item labels (Список меток пунктов) устанавливать не нужно. Кроме того, в выводе будут представлены результаты расчета шкалы альфа для каждого пункта в предположении, что он из шкалы удален.

Далее приведен пример анализа надежности с участием 9 переменных, для которых предыдущий анализ выявил наибольшую надежность (a1, a2, a3, a4, a7, a8, a9, a10, a12). Кроме того, мы зададим некоторые дополнительные параметры и снова включим в выводимые результаты метки переменных.

**Шаг 5а** После выполнения шага 4 должно быть открыто диалоговое окно Reliability Analysis (Анализ надежности), показанное па рис. 19.1. Если вы уже успели поработать с этим окном, щелкните на кнопке Reset (Сброс).

1. Щелкните сначала на переменной a1, чтобы выделить ее, а затем — на кнопке со стрелкой, чтобы переместить переменную в список Items (Пункты).
2. Повторите предыдущее действие для переменных a2, a3, a4, a7, ..., a1O и a12. Поскольку переменные a1 . . . a4 в исходном списке расположены последовательно, для их одновременного выделения можно щелкнуть на переменной a1, пажать клавишу Shift и щелкнуть на переменной a4. То же касается переменных a7. . . . a10.
3. Установите флажок List item labels (Список меток пунктов).
4. Щелкните на кнопке Statistics (Статистики), чтобы открыть диалоговое окно Reliability Analysis: Statistics (Анализ надежности: Статистики), показанное на рис. 19.2.
5. В группе Descriptives for (Описательные статистики для) установите флажки Scale (Шкала) и Scale if item deleted (Шкала, если элемент удален), в группе Summaries (Итоги) — флажки Means (Средние), Variances (Дисперсии) и Correlations (Корреляции), в группе Inter-Item (Между пунктами) — флажок Correlations (Корреляции). Щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Reliability Analysis (Анализ надежности).
6. Щелкните па кнопке OK, чтобы открыть окно вывода.

В следующем примере мы выполним ту же процедуру, что и на шаге 5а, однако вместо коэффициента а будем рассчитывать надежность половинного расщепления.

**Шаг 563** После выполнения шага 4 должно быть открыто диалоговое окно Reliability Analysis (Анализ надежности), показанное на рис. 19.1. Если вы уже успели поработать с этим окном, щелкните на кнопке Reset (Сброс).

1. Щелкните сначала па переменной a1, чтобы выделить ее, а затем — на кнопке со стрелкой, чтобы переместить переменную в список Items (Пункты).
2. Повторите предыдущее действие для переменных a2, a3, a4, a7. . . . a1O и a12. Поскольку переменные a1, ..., a4 в исходном списке расположены последовательно, для их одновременного выделения можно щелкнуть на переменной a1, нажать клавишу Shift и щелкнуть на переменной a4. То же касается переменных a7, ..., a10.
3. Раскройте список Model (Модель) и выберите пункт Split-half (Половинное расщепление), после чего установите флажок List item labels (Список меток пунктов).
4. Щелкните на кнопке Statistics (Статистики), чтобы открыть диалоговое окно Reliability Analysis: Statistics (Анализ надежности: Статистики) (см. рис. 19.2).
5. В группе Descriptives for (Описательные статистики для) установите флажки Scale (Шкала) и Scale if item deleted (Шкала, если элемент удален), в группе Summaries (Итоги) — флажки Means (Средние), Variances (Дисперсии) и Correlations (Корреляции), в группе Inter-Item (Между пунктами) — флажок Correlations (Корреляции). Щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Reliability Analysis: Statistics (Анализ надежности: Статистики).
6. Щелкните на кнопкеO K, чтобы открыть окно вывода.

После выполнения шага 5 программа автоматически активизирует окно вывода. Для просмотра результатов вы при необходимости можете воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

# Печать результатов и выход из программы

Ниже описана типичная процедура печати результатов статистического анализа (или нескольких анализов). После выполнения шага 5 должно быть открыто окно вывода.

Шаг 6 В окне вывода укажите фрагменты, выводимые на печать (см. раздел «Окно вывода» в главе 2), в меню File (Файл) выберите команду Print (Печать), при необходимости задайте параметры печати и щелкните на кнопке OK.

Последнее, что необходимо сделать после завершения исследования и печати результатов, — это выйти из программы SPSS.

Шаг 7 Для выхода из программы в меню File (Файл) выберите команду Exit (Выход).

Иногда после выполнения команды Exit (Выход) на экране могут появляться небольшие диалоговые окна с вопросом о необходимости сохранения сделанных в файлах изменений и кнопками, описывающими возможные варианты ответа. Для завершения работы просто щелкайте на соответствующих кнопках.

# Представление результатов

В этом разделе представлены фрагменты выводимых данных, генерируемые программой при выполнении шагов 5, 5а и 5б. Обратите внимание, что все таблицы выводятся в простом текстовом формате. Сначала приводится список имен переменных с их метками, затем следует таблица Item-total Statistics, но которой мы выбираем переменные для последующего анализа, далее приведены статистики для сумм переменных и корреляции между элементами, таблица Item-total Statistics для 9 элементов, выбранных из 12 исходных, и, наконец, анализ надежности половинного расщепления.

## Коэффициент альфа с параметрами по умолчанию

Первая из полученных таблиц содержит имена всех переменных, которые участвовали в анализе (рис. 19.3). Во второй таблице для каждой переменной показа-

ны статистические характеристики, позволяющие сравнить ее с совокупностью остальных переменных (рис. 19.4).

```
R E L I A B I L I T Y   A N A L Y S I S   -
S C A L E   (A L P H A)

  1.     A1              великодушный
  2.     A2              гуманный
  3.     A3              человечный
  4.     A4              либеральный
  5.     A5              филантроп
  6.     A6              жалостливый
  7.     A7              участливый
  8.     A8              милосердный
  9.     A9              заботливый
 10.     A10             благосклонный
 11.     A11             ласковый
 12.     A12             добродушный
```

**Рис. 19.3.** Имена всех переменных, участвовавших в анализе

Item-total Statistics

| | Scale Mean if Item Deleted | Scale Variance if Item Deleted | Corrected Item-Total Correlation | Alpha if Item Deleted |
|---|---|---|---|---|
| A1 | 5.6750 | 6.3893 | .6430 | .6286 |
| A2 | 5.7250 | 6.5876 | .5464 | .6437 |
| A3 | 5.6750 | 6.8935 | .4251 | .6630 |
| A4 | 5.7250 | 7.1254 | .3242 | .6782 |
| A5 | 5.8167 | 8.4703 | -.1638 | .7449 |
| A6 | 5.8083 | 7.9714 | .0080 | .7227 |
| A7 | 5.7583 | 6.9243 | .4039 | .6661 |
| A8 | 5.6417 | 6.7361 | .5008 | .6517 |
| A9 | 5.7167 | 6.7594 | .4743 | .6552 |
| A10 | 5.7500 | 6.7269 | .4857 | .6534 |
| A11 | 5.8167 | 8.2350 | -.0839 | .7347 |
| A12 | 5.7333 | 6.6006 | .5401 | .6447 |

```
Reliability  Coefficients
N of Cases -     120.0                 N of Items = 12
Alpha -     .6961
```

**Рис. 19.4.** Статистические характеристики всех переменных

# Терминология, используемая при выводе

Далее дана трактовка терминов, используемых программой в окне вывода и относящихся к вычислению коэффициента а с параметрами по умолчанию.

- Scale mean if item deleted (Среднее шкалы, если элемент удален) — для каждой переменной вычисляется сумма остальных 11 переменных по всем 120 объектам файла.

- Scale variance if item deleted (Дисперсия шкалы, если элемент удален) — дисперсия суммы всех переменных с той переменной, которая удалена.
- Corrected item-total correlation (Скорректированная корреляция элемента и суммы) - коэффициент корреляции переменной с суммой остальных 11 переменных.
- Alpha if item deleted (Коэффициент альфа, если элемент удален) — коэффициент а, полученный при удалении соответствующей переменной из шкалы.

Как можно видеть, для трех переменных коэффициент корреляции между ними и суммой остальных переменных (Corrected item-total correlation) оказывается весьма низким, а для переменных а5 и a11 даже имеет отрицательное значение. Для увеличения коэффициента а (указанного в самом конце) необходимо удалить такие переменные из анализа. После исключения трех переменных был проведен анализ, действительно позволивший найти новое, более высокое значение коэффициента а. Таким образом удалось выявить, что наибольшее значение а принимает в случае, если удалить из шкалы переменные с номерами 5, 6, и 11. Обратите внимание, что подобные переменные не исключаются программой автоматически: существуют причины, по которым иногда требуется сохранять эти переменные в шкале.

## Коэффициент альфа после исключения переменных и задания дополнительных параметров

Приведенные результаты формируются после удаления из шкалы трех упомянутых переменных. Первая таблица содержит основные статистические показатели для шкалы: средние, дисперсии и корреляции (рис. 19.5). Таблица Item-total Statistics по своему содержанию аналогична одноименной таблице, приведенной ранее на рис. 19.4, однако здесь вместо 12 исходных переменных содержатся лишь 9 (рис. 19.6). Как легко видеть, значение коэффициента а увеличилось и составляет 0,8251. Данное значение а больше нельзя увеличить, удалив какую-либо из 9 оставшихся переменных.

```
R E L I A B I L I T Y   A N A L Y S I S   -
S C A L E    (A L P H A)
         N of Cases =          120.0
                                                 N of
Statistics       Mean   Variance   Std Dev  Variables
for Scale       4.9250    8.3893    2.8964          9

Item Means        Mean   Minimum   Maximum     Range   Max/Min  Variance
                 .5472     .5000     .6167     .1167    1.2333     .0015

Item Variances    Mean   Minimum   Maximum     Range   Max/Min  Variance
                 .2485     .2384     .2521     .0137    1.0576     .0000
Inter-item
Correlations      Mean   Minimum   Maximum     Range   Max/Min  Variance
                 .3440     .1336     .6325     .4988    4.7329     .0129
```

Рис. 19.5. Основные статистические показатели для шкалы

```
Item-total Statistics

            Scale          Scale        Corrected
            Mean           Variance     Item-          Alpha
            if Item        if Item      Total          if Item
            Deleted        Deleted      Correlation    Deleted

A1          4.3417         6.4621       .6683          .7907
A2          4.3917         6.4924       .6447          .7933
A3          4.3417         6.9495       .4577          .8154
A4          4.3917         7.0806       .3967          .8224
A7          4.4250         6.9859       .4338          .8182
A8          4.3083         6.8369       .5146          .8089
A9          4.3833         6.8938       .4739          .8136
A10         4.4167         6.7157       .5463          .8051
A12         4.4000         6.5445       .6210          .7962

Reliability Coefficients
N of Cases =    120.0                N of Items =  9
Alpha =       .8251
```

**Рис. 19.6.** Статистические характеристики после удаления трех переменных

Ниже дана трактовка новых терминов, используемых программой в окне вывода и относящихся к вычислению коэффициента а после исключения нескольких переменных и задания дополнительных параметров.

- Statistics for scale (Статистики для шкалы) — основные статистические показатели для суммы всех 9 задействованных в шкале переменных.
- Item means (Средние элементов) — описательная информация о средних значениях 9 пунктов шкалы. Среднее значение полученных средних значений равно 0,5472, наименьшее из 9 средних равно 0,5 и т. д.
- Item variances (Дисперсии элементов) — статистика, аналогичная средним элементов, но полученная для их дисперсий.
- Inter-item correlations (Корреляция между элементами) — описательная статистика, характеризующая корреляцию между каждой переменной шкалы и суммой остальных переменных. Для каждой переменной вычисляется коэффициент корреляции, а в выводе приводятся среднее значение, минимум и т. д. полученных коэффициентов. Среднее значение коэффициентов корреляции является величиной $r$ в формуле расчета а.
- Squared multiple correlation (Квадрат множественной корреляции) — представленные в этой секции величины вычислены с помощью уравнения множественной регрессии, позволяющего получать прогнозируемый коэффициент корреляции данной переменной с остальными переменными, участвующими в анализе.
- Alpha (Альфа) — коэффициент а, вычисляемый по формуле, приведенной в начале главы. Число элементов шкалы ($k$) в данном случае равно 9. Поскольку а зависит от числа элементов шкалы, нельзя четко определить критерий

приемлемости полученного значения a. Для большинства случаев можно придерживаться следующих рекомендаций по оценке внутренней согласованности шкалы:

- a > 0,9 — отличная;
- a > 0,8 — хорошая;
- a > 0,7 — приемлемая;
- a > 0,6 — сомнительная;
- a > 0,5 — малопригодная;
- a < 0,5 — недопустимая.

## Надежность половинного расщепления

На рис. 19.7 и 19.8 приведены результаты анализа надежности половинного расщепления с использованием тех же 9 переменных, что и в предыдущем случае. Как правило, такое количество переменных слишком мало для подобного анализа; тем не менее результаты вполне пригодны для интерпретации. Большая часть полученной информации по смыслу аналогична результатам, полученным для коэффициента a. Мы приводим лишь те фрагменты, которые различны для этих двух типов анализа.

```
R E L I A B I L I T Y   A N A L Y S I S  -
S C A L E   (S P L I T)
          N of Cases =        120.0
                                             N of
Statistics for   Mean Variance   Std Dev   Variables
       Part 1  2 .7333   2.9031   1.7038          5
       Part 2  2 .1917   2.0722   1.4395          4
       Scale   4 .9250   8.3893   2.8964          9

Item Means       Hean   Minimum   Maximum    Range   Max/Min  Variance
       Part 1   .5467    .5000     .5833    .0833   1.1667     .0013
       Part 2   .5479    .5083     .6167    .1083   1.2131     .0023
       Scale    .5472    .5000     .6167    .1167   1.2333     .0015

Item Variances   Mean   Minimum   Maximum    Range   Max/Min  Variance
       Part 1   .2489    .2451     .2521    .0070   1.0286     .0000
       Part 2   .2481    .2384     .2520    .0137   1.0573     .0000
       Scale    .2485    .2384     .2521    .0137   1.0576     .0000

Inter-item
Correlations     Mean   Minimum   Maximum    Range   Max/Min  Variance
       Part 1   .3339    .1336     .6325    .4988   4.7329     .0239
       Part 2   .3624    .2328     .5652    .3324   2.4278     .0145
       Scale    .3440    .1336     .6325    .4988   4.7329     .0129
```

Рис. 19.7. Основные статистические показатели для шкалы при анализе надежности половинного расщепления

```
R E L I A B I L I T Y   A N A L Y S I S  -
S C A L E   (S P L I T)
Reliability Coefficients    9 items

Correlation between
forms               = .6960  Equal-length Spearman-Brown =  .8207

Guttman Split-half = .8139  Unequal-length Spearman-Brown =.8223

Alpha for part 1   = .7143  Alpha for part 2 =              .6949

  5 items in part 1               4 items in part 2
```

Рис. 19.8. Некоторые статистические характеристики при анализе надежности половинного расщепления

Секции Item Means, Item Variances и Inter-Item correlations содержат по 3 строки, первые две из которых представляют две половины шкалы, а третья — шкалу в целом. Как можно заметить, третьи строки каждой секции этой таблицы содержат те *же* значения, что и соответствующие секции таблицы, которая была получена ранее для коэффициента а (см. рис. 19.5). Обратите внимание, что значения а, вычисленные для каждой половины шкалы, ниже, чем для шкалы в целом. Это отражает тот факт, что при уменьшении числа элементов шкалы ее внутренняя согласованность снижается.

Ниже дана трактовка нескольких новых терминов, относящихся к вычислению надежности половинного расщепления.

- Correlation between forms (Корреляция между формами) — приближенное значение надежности шкалы, рассчитанное в предположении, что она содержит 5 элементов.
- Equal-length Spearman-Brown (Коэффициент эквивалентных форм Спирмена—Брауна) — результат гипотетического анализа надежности половинного расщепления шкалы, содержащей 10 элементов.
- Guttman Split-Half (Критерий надежности половинного расщепления Гуттмана) - значение"Надежности, полученное путем вычисления нижних пределов.
- Unequal-length Spearman-Brown (Критерий неэквивалентных форм Спирмена—Брауна) — результат гипотетического анализа надежности половинного расщепления шкалы, содержащей не 10, а 9 элементов, что приводит к необходимости формирования двух неравных групп.

# 20 Факторный анализ



В последние 30-40 лет факторный анализ приобрел значительную популярность в психологических и социальных исследованиях. Во многом этому способствовала разработка Раймондом Кеттеллем (Raymond B. Cattell) знаменитого 16-факторного личностного опросника (16PF). Именно при помощи факторного анализа ему удалось свести около 4500 наименований личностных особенностей к 187 вопросам, которые, в свою очередь, позволяют измерить 16 различных свойств личности. Факторный анализ дает возможность количественно определить нечто непосредственно не измеряемое, исходя из нескольких доступных измерению переменных. Например, характеристики «посещает развлекательные мероприятия», «много разговаривает», «охотно идет на контакт с любым незнакомым человеком» могут служить оценками качества «общительность», которое непосредственно не поддается количественному измерению. Факторный анализ позволяет установить для большого числа исходных признаков сравнительно узкий набор «свойств», характеризующих связь между группами этих признаков и называемых факторами. Процедура факторного анализа состоит из четырех основных стадий:

1. Вычисление корреляционной матрицы для всех переменных, участвующих в анализе.
2. Извлечение факторов.
3. Вращение факторов для создания упрощенной структуры.
4. Интерпретация факторов.

Первые три операции мы кратко рассмотрим в следующих нескольких разделах; последняя операция описывается лишь на концептуальном уровне.

Как и многие другие методы, рассматриваемые в этой книге, факторный анализ является весьма сложной процедурой. И нескольких страниц краткого обзора явно недостаточно для его изучения и исчерпывающего понимания. Поэтому желательно, чтобы перед проведением факторного анализа у вас уже были более глубокие знания о нем.

# Вычисление корреляционной матрицы

Первая операция, которая производится при выполнении факторного анализа, — это вычисление корреляционной матрицы для переменных, участвующих в анализе. Уже по наличию подобного начального действия можно сделать вывод о том, что факторный анализ основан на взаимодействии переменных. Для проведения факторного анализа совсем не обязательно специально строить корреляционную матрицу: при необходимости программа SPSS создаст ее сама на основе данных файла. Иногда не нужны даже исходные данные — достаточно иметь корреляционную матрицу, которая в этом случае вводится в командный файл SPSS. Однако подобный подход весьма сложен и в рамках этой книги не рассматривается. При необходимости вы можете обратиться к руководству пользователя программы SPSS.

# Извлечение факторов

Извлечение факторов является следующим этапом факторного анализа. С математической точки зрения извлечение факторов имеет определенную аналогию с множественным регрессионным анализом. Если вы обратитесь к главе 18, то вспомните, что первым шагом в множественном регрессионном анализе является выбор той независимой переменной, которая обусловливает наибольшую долю дисперсии зависимой переменной. Затем операция повторяется для оставшихся независимых переменных до тех пор, пока доля дисперсии не перестанет быть значимой. В факторном анализе существует аналогичная процедура, и для ее описания достаточно лишь немного изменить соответствующий раздел главы 18.

Извлечение фактора начинается с подсчета суммарного разброса значений всех участвующих в анализе переменных (данная величина чем-то похожа на общую сумму квадратов). Для этого «суммарного разброса» непросто подобрать логическую интерпретацию, однако он является вполне строго определенной математической величиной. Первой задачей факторного анализа является выбор взаимодействующих переменных, чья взаимная корреляция обусловливает наибольшую долю общей дисперсии. Эти переменные образуют *первый фактор.* Затем первый фактор исключается, и из оставшегося множества переменных снова выбираются те, чье взаимодействие определяет наибольшую долю оставшейся общей дисперсии. Эти переменные образуют *второй фактор.* Процедура извлечения факторов продолжается до тех пор, пока не будет исчерпана вся общая дисперсия переменных.

По умолчанию в процедуре факторного анализа каждая переменная имеет единичное значение *общности.* Этот показатель равен доле дисперсии переменной,

обусловленной совокупным влиянием факторов. Общность можно сравнить с множественным коэффициентом корреляции *R,* принимающим значение 0 в случае, если факторы не влияют на переменную, и значение 1 в случае, если дисперсия переменной целиком определяется выделяемыми факторами. Перед началом извлечения факторов единичное значение общности установлено по умолчанию для всех переменных, участвующих в факторном анализе.

После того как процедура извлекает первый фактор, напротив его номера появляется его *собственное значение,* например рядом с числом 1 появляется значение 5,13312. Собственное значение фактора пропорционально доле общей дисперсии, определяемой данным фактором (обратите внимание, что здесь речь идет о факторе, а не о переменной, как было в случае общности). Первое собственное значение всегда является наибольшим и превышает единицу; это определяется алгоритмом работы процедуры, согласно которому факторы извлекаются в порядке убывания их влияния на дисперсию переменных. Затем вычисляется процент дисперсии, обусловливаемый данным фактором и равный отношению собственного значения фактора к числу переменных, а также соответствующий кумулятивный (накопленный) процент. С извлечением каждого нового фактора собственные значения уменьшаются, а кумулятивный процент приближается к 100. Обратите внимание, что в контексте процедуры извлечения факторов ни разу не прозвучал термин «значимость»: в отличие от регрессионного анализа извлечение факторов происходит до тех пор, пока не будет исчерпана вся дисперсия переменных, независимо от того, является ли значимым влияние фактора на дисперсию или нет.

# Выбор и вращение факторов

За очень редкими исключениями для исследователя не представляют интереса все извлеченные факторы. Если факторов окажется столько же, сколько исходных переменных, факторный анализ теряет смысл, поскольку его целью является сокращение исходного набора переменных.

Итак, нужно принять решение, какие из факторов следует оставить для дальнейшего анализа. Здесь, в первую очередь, рекомендуется руководствоваться здравым смыслом и оставлять те факторы, которые имеют понятную теоретическую или логическую интерпретацию. Однако не всегда представляется возможным заранее установить назначение каждого фактора, и поэтому исследователи на первом этапе обычно используют формальные критерии. По умолчанию при выполнении команды Factor (Факторный анализ) все факторы, чьи собственные значения превышают единицу, сохраняются для дальнейшего анализа. Поскольку число факторов равно числу переменных, лишь для небольшого количества факторов собственные значения оказываются больше единицы, значит выполнение команды с параметрами по умолчанию позволяет радикально сократить число факторов. Существуют и другие критерии выделения факторов (например, критерий «каменистой осыпи» Р. Кеттелла); кроме того, вы можете выбирать факторы, основываясь на известных вам особенностях конкретного файла данных. В любом случае окончательное решение о числе факторов обычно принимается после ин-

терпретации факторов, следовательно, факторный анализ предполагает неоднократное выделение различного числа факторов. В разделе пошаговых процедур рассмотрены несколько вариантов выполнения факторного анализа, отличные от принятого но умолчанию.

Следующим шагом после выделения факторов является их вращение. Вращение требуется потому, что изначально структура факторов, будучи математически корректной, как правило, трудна для интерпретации. Целью вращения является получение простой структуры, которой соответствует большое значение нагрузки каждой переменной только по одному фактору и малое по всем остальным факторам. *Нагрузка* отражает связь между переменной и фактором, являясь подобием коэффициента корреляции. Значение нагрузки лежит в пределах от -1 до 1. Идеальная простая структура предполагает, что каждая переменная имеет нулевые значения нагрузок для всех факторов кроме одного, для которого нагрузка этой нереременной близка к 1 (−1). Графическая интерпретация процедуры вращения представлена на рис. 20.1. До вращения (слева) звездочки, соответствующие переменным, расположены на удалении от осей факторов. После поворота осей (справа) переменные оказываются вблизи осей, что соответствует максимальной нагрузке каждой переменной только по одному фактору. На практике строгая ориентация переменных вдоль осей факторов обычно не достигается, однако операция поворота позволяет приблизиться к желательному результату. После выполнения вращения факторов программа SPSS включает в вывод похожие графические изображения, для удобства снабдив их таблицей координат.

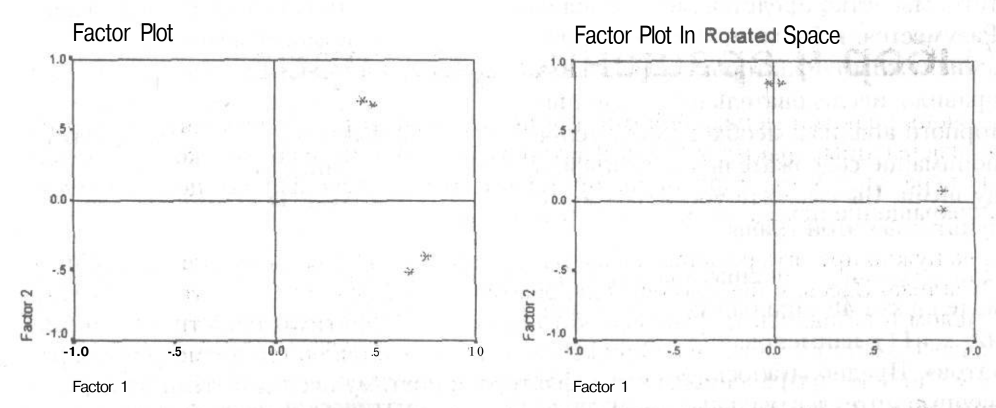



**Рис. 20.1.** Графическая интерпретация процедуры вращения факторов

Вращение факторов нисколько не влияет на математическую строгость анализа: взаимное положение звездочек-переменных не меняется при повороте осей. Изначально процедура вращения выполнялась вручную, и исследователь самостоятельно задавал такое положение осей, при котором точки переменных максимально бы к ним приближались. Теперь SPSS позволяет выполнить несколько вариантов вращения, поворачивающих оси так, чтобы получить простую структуру, удовлетворяющую тому или иному критерию. Наиболее популярным вариантом вращения является метод Varimax.

Вариант вращения Varimax является ортогональным, поскольку при таком вращении оси сохраняют свое взаимное расположение иод прямым углом. Иногда можно получить более предпочтительную простую структуру, если изменить угол между осями. Для этого предназначены варианты вращения Direct Oblimin и Promax. Наличие непрямого угла между осями факторов означает, что они не являются *полностью* независимыми друг от друга. Поскольку в реальных исследованиях факторы действительно не являются абсолютно независимыми, отклонение угла между осями от прямого при вращении вполне допустимо. Неортогональное вращение — достаточно сложная процедура, и для ее применения необходимо четко представлять себе суть происходящего. В целом аналогичную рекомендацию можно дать относительно факторного анализа вообще: не следует применять его без предварительного изучения соответствующих разделов статистики. Варианты вращения Direct Oblimin и Promax описаны в разделе пошаговых процедур, однако мы не демонстрируем их результаты в разделе «Представление результатов» ввиду сложности последних.

# Интерпретация факторов

Итак, пусть в некоторой ситуации (близкой к идеальной) путем вращения мы добились того, что значение нагрузки для рассматриваемого фактора является большим (более 0,5), а для остальных факторов — малым (менее 0,2); кроме того, мы четко представляем смысл нашего фактора, то есть то, что он измеряет. Разумеется, в большинстве исследований переменные могут взаимодействовать с «ненужным» фактором, а нередко таких факторов может быть несколько. Как *правило*, исследователь не ограничивается только числовыми результатами факторного анализа; необходимым условием успеха факторного анализа является понимание содержательной специфики конкретных данных и взаимосвязей между ними. Вы сможете убедиться в этом при чтении раздела «Представление результатов» этой главы.

Для факторного анализа мы будем использовать данные реального тестирования интеллекта 46 школьников. Тест включал в себя 11 субтестов (переменные и1, и2, ..., и11), наименования которых вы найдете в разделе «Представление результатов». Предполагалось, что эти 11 субтестов позволят измерить 3 и более обобщенные интеллектуальные характеристики: математические, вербальные и невербальные (образные). Факторный анализ должен был установить соотношение субтестов и факторов.

Файл данных, который мы будем использовать в примере, называется TestIQ.sav; число объектов (*N*) равно 46.

В разделе пошаговых процедур приведено два варианта шага 5. Шаг 5 иллюстрирует простейший вариант факторного анализа, в котором используются значения по умолчанию для всех параметров. Шаг 5а, напротив, включает многие из действий, упомянутых ранее в этой главе.

# Пошаговые алгоритмы вычислений

При проведении факторного анализа сначала выполняются три подготовительных шага.

**Шаг 1** Создайте новый файл данных или подготовьте существующий. Все необходимые действия описаны в главе 3.

**Шаг 2** Запустите программу SPSS при помощи значка на рабочем столе или команды Пуск ▸ Программы ▸ SPSS for Windows ▸ SPSS 11.5 for Windows (Start ▸ Programs ▸ SPSS for Windows ▸ SPSS 11.5 for Windows) главного меню Windows. В открывшемся после запуска программы диалоговом окне SPSS for Windows щелкните на кнопке Cancel (Отмена).

После выполнения этого шага на экране появится окно редактора данных SPSS.

**Шаг 3** Откройте файл данных, с которым вы намерены работать (в нашем случае это файл TestlQ.sav). Если он расположен в текущей папке, то выполните следующие действия:

1. Выберите в меню File (Файл) команду Open ▸ Data (Открытие ▸ Данные) или щелкните на кнопке Open File (Открытие файла) панели инструментов.
2. В открывшемся диалоговом окне дважды щелкните на имени TestlQ.sav или введите его с клавиатуры и щелкните на кнопке OK.

Независимо от того, открыта программа SPSS только что или какие-то процедуры уже выполнялись, в верхней части главного окна должна присутствовать строка меню (она показана ниже). Пока строка меню присутствует на экране, доступны все команды анализа данных. При этом окно с данными видеть не обязательно.




При работе со сводными таблицами или при редактировании диаграмм некоторые пункты строки меню могут исчезать или меняться. Чтобы вернуться к главному окну и стандартной строке меню, щелкните на кнопке свертывания или восстановления текущего окна.

После завершения шага 3 на экране должно присутствовать окно редактора данных.

**Шаг 4** В меню Analyze (Анализ) выберите команду Data Reduction ▸ Factor (Сокращение данных ▸ Факторный анализ). На экране появится диалоговое окно Factor Analysis (Факторный анализ), показанное на рис. 20.2.

Диалоговое окно Factor Analysis (Факторный анализ) почти ничем не отличается от большинства диалоговых окоп команд статистических операций SPSS: список доступных переменных находится в левой части, список Variables (Переменные), предназначенный для указания имен переменных, участвующих в анализе, —

в центре, а пять кнопок, с помощью которых задаются параметры анализа, — в правой. Несмотря на то что существует множество математических нюансов, касающихся факторного анализа, для его проведения в простейшем случае достаточно всего лишь поместить в список Variables (Переменные) несколько переменных, выбрать вариант вращения и щелкнуть на кнопке OK. Далее будет приведен пример, в котором мы выберем 11 переменных (и1, ..., и11) и проведем факторный анализ с параметрами по умолчанию. Кнопки, расположенные в нижней части окна, позволяют управлять множеством разнообразных параметров анализа и могут использоваться независимо друг от друга в любом порядке.

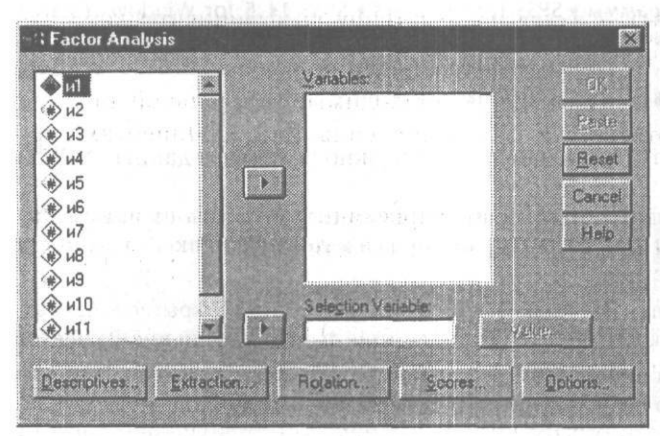



Рис. 20.2. Диалоговое окно Factor Analysis

Кнопка Descriptives (Описательные статистики) предназначена для открытия диалогового окна Factor Analysis: Descriptives (Факторный анализ: Описательные статистики), представленного на рис. 20.3.

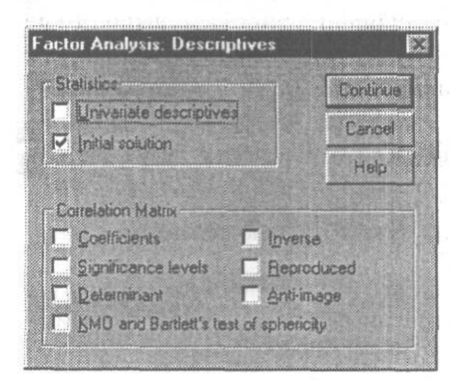



Рис. 20.3. Диалоговое окно Factor Analysis: Descriptives

В этом окне находятся две группы флажков: Statistics (Статистики) и Correlation Matrix (Корреляционная матрица). Флажок Univariate descriptives (Одномерные описательные статистики) позволяет получить в таблице с четырьмя столбцами имена

переменных, их средние значения, стандартные отклонения и метки. Мы нередко в процессе интерпретации вывода будем обращаться к значениям, содержащимся в этой таблице. Флажок Initial Solution (Начальное решение) по умолчанию установлен и отвечает за включение в выводимые данные имен переменных, начальных общностей (по умолчанию равных 1,0), факторов, собственных значений, а также общего и кумулятивного процентов общей дисперсии для каждого фактора. В группу Correlation Matrix (Корреляционная матрица) входит 7 флажков, управляющих выводом корреляционной матрицы; ниже приведены описания четырех из них, которые используются чаще других.

- Coefficients (Коэффициенты) — включает в матрицу коэффициенты корреляции, ради которых она обычно создается.
- Significance levels (Уровни значимости) — приводимые в отдельной таблице значения $p$-уровней, соответствующих вычисленным коэффициентам корреляции.
- Determinant (Детерминант) — детерминант корреляционной матрицы, использующийся для критериев многомерной нормальности.
- KMO and Barlett test of sphericity (Критерии KMO и сферичности Барлетта) — два критерия: на многомерную нормальность (Барлетта) и адекватность выборки (KMO определяет применимость факторного анализа к выбранным переменным). По умолчанию эти тесты не проводятся, но они обеспечивают процедуру важной начальной статистической информацией. Более подробно данный параметр рассматривается в разделе «Представление результатов».

Щелчок па кнопке Extraction (Извлечение) в окне Factor Analysis (Факторный анализ) приводит к открытию диалогового окна Factor Analysis: Extraction (Факторный анализ: Извлечение), представленного на рис. 20.4 и позволяющего выбрать метод извлечения, задать критерий числа факторов, сконфигурировать выводимые данные, связанные с процедурой извлечения, а также определить максимальное количество итераций, приводящих к получению результата.

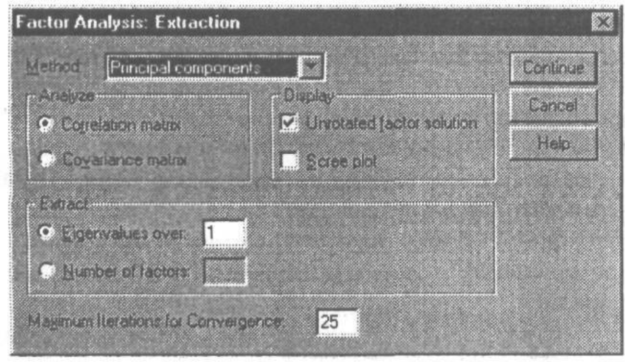



**Рис. 20.4.** Диалоговое окно Factor Analysis: Extraction

Раскрывающийся список Method (Метод) содержит 7 пунктов. По умолчанию выбран пункт Principal components (Главные компоненты); остальные пункты, соответствующие методам извлечения, перечислены ниже:

- Unweighted least squares (Метод невзвешенных наименьших квадратов);
- Celeralized least squares (Обобщенный метод наименьшихквадратов);
- Maximum likehood (Метод максимального правдоподобия);
- Principal-axis factoring (Метод главных факторов);
- Alpha-factoring (Альфа-факторный анализ);
- Image factoring (Факторный анализ образов).

Формально именно эти 6 перечисленных методов относятся к факторному анализу. Тем не менее мы ограничимся анализом главных компонентов, как установленного по умолчанию и к тому же наиболее простого и чаще всего используемого метода.

В группе Analyze (Анализ) с помощью переключателей Correlation matrix (Корреляционная матрица) и Covariance matrix (Ковариационная матрица) можно выбрать начальные условия анализа. Переключатели группы Extract (Извлечение) позволяют задать критерий числа извлекаемых факторов: если установлен переключатель Eigenvalues over (Собственные значения больше), то можно воспользоваться единичным значением, установленным по умолчанию, либо задать свое значение. Если установлен переключатель Number of factors (Число факторов ), можно извлечь заданное число факторов, которое указывается в поле рядом с переключателем. Группа Display (Отображать) содержит два флажка: Unrotated factor solution (Факторное решение до вращения), установленный по умолчанию, и Scree Plot (График собственных значений), часто используемый исследователями для определения числа факторов по критерию Р. Кеттелла. Если вы достаточно опытны в факторном анализе, то, вероятно, предпочтете установить оба указанных флажка. Наконец, при желании вы можете изменить значение в поле Maximum Iterations for Convergence (Максимальное число итераций для сходимости), по умолчанию установленное равным 25.

Щелчок на кнопке Rotation (Вращение) в окне Factor Analysis (Факторный анализ) открывает диалоговое окно Factor Analysis: Rotation (Факторный анализ: Вращение), представленное на рис. 20.5. При выполнении факторного анализа можно выбрать один из трех ортогональных методов вращения: наиболее популярные методы Varimax, Equamax и Quartimax. Переключатели Direct Oblimin и Promax позволяют выполнять неортогональное вращение факторов. Значения в полях Delta (Дельта) и Карра (Каппа) практически всегда рекомендуется оставлять установленными по умолчанию. Как говорилось ранее, настоятельно не рекомендуется использовать неортогональное вращение без исчерпывающих теоретических знаний в области факторного анализа.

Флажок Rotated solution (Факторы после вращения) в группе Display (Отображать) по умолчанию установлен, если выбран один из методов вращения. Он позволяет

включить в выводимые данные наиболее важные результаты факторного анализа. Более подробно об этом рассказано в разделе «Представление результатов». Если вам необходим визуальный анализ структуры факторов после вращения, установите флажок Loading plot(s) (Диаграммы нагрузок). По умолчанию в выводимые данные будет включено трехмерное изображение первых трех факторов, если факторов больше двух. Вы можете перемещать оси и менять точку наблюдения изображения, но, пожалуй, неизменным в большинстве случаев будет одно: вы не сможете уловить смысла того, что видите. Однако из созданного трехмерного изображения можно попытаться получить двухмерный срез наподобие представленного на рис. 20.6. Для этого нужно активизировать окно редактора диаграмм, дважды щелкнув на изображении, в меню Gallery (Галерея) выбрать команду Scatter (Разброс), в появившемся диалоговом окне в качестве типа изображения выбрать вариант Simple (Простое) и, наконец, щелкнуть иа кнопке Replace (Заменить), после чего можно приступать к собственно редактированию.

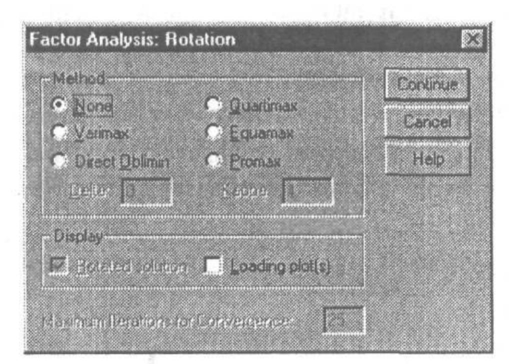



**Рис. 20.5.** Диалоговое окно Factor Analysis: Rotation

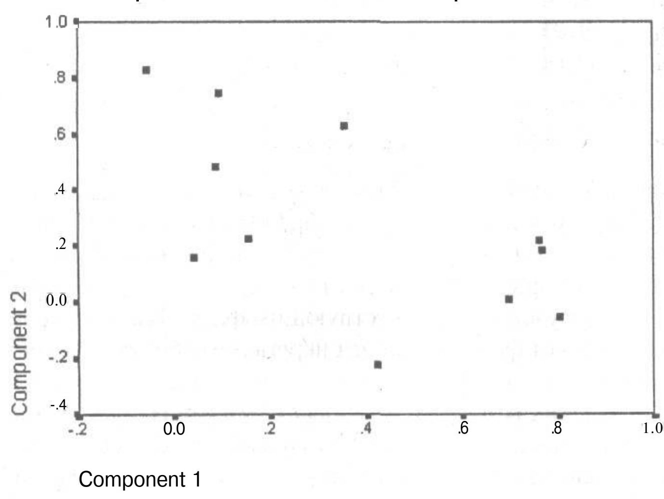



Рис. **20.6.** График двух первых факторов после вращения

Двухмерные изображения поддаются интерпретации гораздо легче, в особенности если вы обратитесь к величинам, содержащимся в факторной матрице (см. раздел «Представление результатов»). Обратите внимание, что горизонтальной осью диаграммы является ось Component 1 (Компонент 1), соответствующая математическим способностям, а вертикальной осью — ось Component 2 (Компонент 2), соответствующая вербальным способностям. Наличие 4 точек в правой части диаграммы свидетельствует о том, что четыре переменные имеют высокие нагрузки по фактору 1 (0,6-0,8) и низкие нагрузки по фактору 2 (−0,1−0,4).

Вернемся к основному диалоговому окну команды Factor (Факторный анализ) и обратимся к кнопке Scores (Величины). При щелчке на ней открывается диалоговое окно, позволяющее сохранить статистические величины как переменные. Мы не приводим рисунок данного диалогового окна; вам достаточно знать, что для включения в выводимые данные матрицы факторных коэффициентов необходимо установить флажок Display factor score coefficient matrix (Отобразить матрицу факторных коэффициентов). Установка флажка Save as variables (Сохранить как переменные) позволяет сохранить в файле данных вычисленные значения факторов для объектов в качестве новых неременных.

Кнопка Options (Параметры) в окне Factor Analysis (Факторный анализ) предназначена для вызова диалогового окна Factor Analysis: Options (Факторный анализ: Параметры), представленного на рис. 20.7.

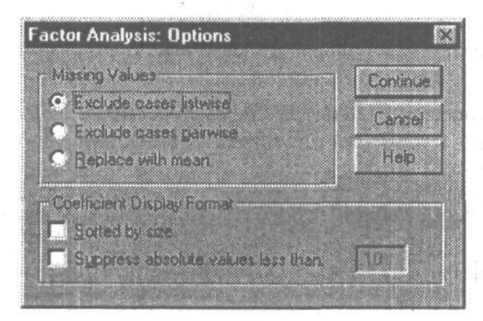



**Рис. 20.7.** Диалоговое окно Factor Analysis: Options

Элементы интерфейса, расположенные в этом окне в группе Coefficient Display Format (Формат вывода коэффициентов), управляют отображением матрицы факторных нагрузок после вращения. Часто полезно установить флажок Sorted by size (Сортировать по величине), который позволяет отсортировать переменные в соответствии с величиной их иафузок но соответствующим факторам. Так, если 6 переменных максимально нагружают фактор 1, то эти переменные будут перечислены в порядке убывания их нагрузок в столбце с названием Factor 1 (Фактор 1). Аналогичным образом будут выведены нагрузки для всех остальных факторов. В разделе «Представление результатов» вы можете ознакомиться с видом этой таблицы. Флажок Suppress absolute values less than (Отсечь абсолютные величины, меньшие) позволяет исключить из таблицы все значения загрузок, меньшие

заданной величины. Это делается для того, чтобы исследователь в окне вывода мог избавиться от тех величин нагрузок, которые не имеют для него значения. Группа переключателей Missing Values (Пропущенные значения) позволяет выбрать режим обработки отсутствующих значений (см. главу 4).

Далее приведены два варианта выполнения шага 5. В первом из них почти для всех параметров факторного анализа оставлены значения по умолчанию; второй вариант является более интересным и иллюстрирует применение некоторых из упоминавшихся параметров.

В следующем примере иллюстрируется применение факторного анализа к 11 переменным и1, ..., и11 файла TestIQ.sav с параметрами по умолчанию и вращением по методу Varimax.

**Шаг 5** После выполнения предыдущего шага у вас должно быть открыто диалоговое окно Factor Analysis (Факторный анализ), показанное на рис. 20.2. Если вы уже успели поработать с этим окном, щелкните на кнопке Reset (Сброс).

1. Щелкните на переменной и1, нажмите клавишу Shift и, не отпуская ее, щелкните на переменной и11. В результате окажутся выделенными все промежуточные переменные, начиная от переменной и1 и заканчивая переменной и11.
2. Щелкните на верхней кнопке со стрелкой, чтобы переместить выделенные переменные в список Variables (Переменные).
3. Щелкните на кнопке Rotation (Вращение), чтобы открыть диалоговое окно Factor Analysis: Rotation (Факторный анализ: Вращение), показанное на рис. 20.5.
4. В группе Method (Метод) установите переключатель Varimax и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Factor Analysis (Факторный анализ).
5. Щелкните на кнопке ОК, чтобы открыть окно вывода.

При выполнении этого шага проводится факторный анализ, включающий следующие операции:

1. Вычисление корреляционной матрицы для 11 переменных, задействованных в анализе.
2. Извлечение 11 факторов методом главных компонентов.
3. Выбор для вращения всех факторов, чьи собственные значения не меньше 1.
4. Вращение факторов по методу Varimax.
5. Вывод матрицы преобразования факторов.

В следующем примере проводится факторный анализ с участием тех же 11 переменных, что и в предыдущем случае, однако теперь задаются некоторые дополнительные параметры. Так, мы включим в вывод одномерные описательные статистики всех переменных, коэффициенты корреляции, а также применим

критерии многомерной нормальности и адекватности выборки. Для извлечения факторов будет использоваться метод главных компонентов, а для отображения — график собственных значений. Вращение факторов будет производиться методом Varimax, а благодаря установке флажка Loading plot(s) (Диаграммы нагрузок), мы отобразим факторную структуру после вращения. Наконец, мы отсортируем переменные по величине их нагрузок по факторам.

**Шаг 5а** После выполнения шага 4 у вас должно быть открыто диалоговое окно Factor Analysis (Факторный анализ), показанное на рис. 20.2. Если вы уже успели поработать с этим окном, щелкните па кнопке Reset (Сброс).

1. Щелкните на переменной и1, нажмите клавишу Shift и, не отпуская ее, щелкните на переменной и11. В результате окажутся выделенными все промежуточные переменные, начиная от переменной и1 и заканчивая переменной и11.
2. Щелкните на верхней кнопке со стрелкой, чтобы переместить выделенные переменные в список Variables (Переменные).
3. Щелкните на кнопке Descriptives (Описательные статистики), чтобы открыть диалоговое окно Factor Analysis: Descriptives (Факторный анализ: Описательные статистики), показанное на рис. 20.3.
4. В группе Statistics (Статистики) установите флажок Univariate Descriptives (Одномерные описательные статистики), в группе Correlation matrix (Корреляционная матрица) — флажки Coefficients (Коэффициенты) и KMO and Barlett's Test of sphericity (Критерии KMO и сферичности Барлетта) и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Factor Analysis (Факторный анализ).
5. Щелкните на кнопке Extraction (Извлечение), чтобы открыть диалоговое окно Factor Analysis: Extraction (Факторный анализ: Извлечение), показанное на рис. 20.4.
6. Установите флажок Scree Plot (График собственных значений) и щелкните па кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Factor Analysis (Факторный анализ).
7. Щелкните па кнопке Rotation (Вращение), чтобы открыть диалоговое окно Factor Analysis: Rotation (Факторный анализ: Вращение), показанное на рис. 20.5.
8. В группе Method (Метод) установите переключатель Varimax, в группе Display (Отображать) — флажок Loading plot(s) (Диаграммы нагрузок) и щелкните па кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Factor Analysis (Факторный анализ).
9. Щелкните на кнопке Options (Параметры), чтобы открыть диалоговое окно Factor Analysis: Options (Факторный анализ: Параметры), показанное на рис. 20.7.
10. Установите флажок Sorted by size (Сортировать по величине) и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Factor Analysis (Факторный анализ).

И. Щелкните на кнопке OK, чтобы открыть окно вывода.

После выполнения шага 5 программа автоматически активизирует окно вывода. Для просмотра результатов вы при необходимости можете воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

# Печать результатов и выход из программы

Ниже описана типичная процедура печати результатов статистического анализа (или нескольких анализов). После выполнения шага 5 должно быть открыто окно вывода.

**Шаг 6** В окне вывода укажите фрагменты, выводимые на печать (см. раздел «Окно вывода» в главе 2), в меню File (Файл) выберите команду Print (Печать), при необходимости задайте параметры печати и щелкните на кнопке OK.

Последнее, что необходимо сделать после завершения исследования и печати результатов, — это выйти из программы SPSS.

**Шаг 7** Для выхода из программы в меню File (Файл) выберите команду Exit (Выход).

Иногда после выполнения команды Exit (Выход) на экране могут появляться небольшие диалоговые окна с вопросом о необходимости сохранения сделанных в файлах изменений и кнопками, описывающими возможные варианты ответа. Для завершения работы просто щелкайте на соответствующих кнопках.

# Представление результатов

Далее приведены данные, сгенерированные программой при выполнении шага 5а. Сначала мы рассмотрим фрагмент, полученный благодаря установке флажка KMO and Barlett's Test of sphericity (Критерии KMO и сферичности Барлетта) в окне Factor Analysis: Descriptives (Факторный анализ: Описательные статистики). Этот фрагмент показан на рис. 20.8.

KMO and Bartlett's Test

| | Bartletfs Test of Sphericity | | |
|---|---|---|---|
| Kaiser-Meyer-Olkin Measure of Sampling Adequacy. | Approx. Chi-Square | df | Sig. |
| .704 | 146.069 | 55 | .000 |

**Рис. 20.8.** Результаты тестов KMO и сферичности Барлетта

# Терминология, используемая при выводе

Ниже дана трактовка терминов, используемых программой в окне вывода и относящихся к критериям КМО и сферичности Барлетта.

- Kaiser—Meyer—Olkin Measure of Sampling Adequacy (Критерий адекватности выборки Кайзера—Мейера—Олкина) — величина, характеризующая степень применимости факторного анализа к данной выборке:
  - более 0,9 — безусловная адекватность;
  - более 0,8 — высокая адекватность;
  - более 0,7 — приемлемая адекватность;
  - более 0,6 — удовлетворительная адекватность;
  - более 0,5 — низкая адекватность;
  - менее 0,5 — факторный анализ неприменим к выборке.
- Barlett's Test of Sphericity (Критерий сферичности Барлетта) — критерий многомерной нормальности для распределения переменных. Критерий многомерной нормальности проверяет, отличаются ли корреляции от 0. Значение *p*-уровня, меньшее 0,05, указывает на то, что данные вполне приемлемы для проведения факторного анализа.

В первой из двух таблиц, представленных на рис. 20.9, перечислены имена переменных и общности. Четыре столбца второй таблицы содержат характеристики факторов: их порядковые номера (с 1 по 11), процент общей дисперсии, обусловленной фактором, и соответствующий кумулятивный процент.

Как видите, в первом столбце первой таблицы перечислены метки, а не имена (и1, ..., и11) переменных. Если вы считаете, что метки занимают слишком много места в окне вывода, в меню Edit (Редактирование) выберите команду Options (Параметры), в открывшемся диалоговом окне перейдите на вкладку Output Labels (Отображение меток), в третьем сверху раскрывающемся списке вместо пункта Labels (Метки) выберите пункт Names (Имена) и щелкните на кнопке OK. После этого в таблицах окна вывода всегда будут представлены имена, а не метки переменных. Далее дана трактовка терминов, используемых программой в окне вывода и относящихся к общности и характеристикам факторов.

- Principle Component Analysis (Анализ главных компонентов) — метод извлечения, используемый программой SPSS по умолчанию.
- Initial (Начальные общности) — по умолчанию все переменные имеют исходное значение общности, равное 1.
- Extraction (Конечные общности) — значения общностей после извлечения факторов.
- Component (Компонент) — номер извлеченного фактора.

**Communalities**

| | Initial | Extraction |
|---|---|---|
| осведомленность | 1.000 | .572 |
| скрытые фигуры | 1.000 | .618 |
| пропущенные слова | 1.000 | .562 |
| счет в уме | 1.000 | .687 |
| понятливость | 1.000 | .573 |
| исключение изображений | 1.000 | .502 |
| аналогии | 1.000 | .631 |
| числовые ряды | 1.000 | .628 |
| умозаключения | 1.000 | .565 |
| геометрическое сложение | 1.000 | .707 |
| заучивание слов | 1.000 | .693 |

**Extraction Method: Principal Component Analysis.**

**Total Variance Explained**

| Component | Initial Eigenvalues | | |
|---|---|---|---|
| | Total | % of Variance | Cumulative % |
| 1 | 3.701 | 33.646 | 33.646 |
| 2 | 1.796 | 16.332 | 49.978 |
| 3 | 1.240 | 11.275 | 61.253 |
| 4 | .783 | 7.121 | 68.374 |
| 5 | .763 | 6.933 | 75.307 |
| 6 | .711 | 6.468 | 81.775 |
| 7 | .629 | 5.720 | 87.495 |
| 8 | .439 | 3.995 | 91.490 |
| 9 | .383 | 3.478 | 94.968 |
| 10 | .335 | 3.045 | 98.013 |
| 11 | .219 | 1.987 | 100.000 |

**Extraction Method: Principal Component Analysis.**

**Рис. 20.9. Имена переменных, общности и характеристики факторов**

- Initial Eigenvalues (Исходные собственные значения) — доля дисперсии, обусловленная фактором.
- % of variance (Процент дисперсии) — процент дисперсии, обусловленный фактором, равный отношению собственного значения фактора к сумме исходных общностей (в данном случае равной 11).
- Cumulative % (Кумулятивный процент) — кумулятивный (накопленный) процент дисперсии.

Приведенная на рис. 20.10 диаграмма называется *графиком собственных значений,* или *диаграммой каменистой осыпи* (scree plot). Она представляет точки, соответствующие собственным значениям, в пространстве двух координат. Этот тип диаграммы иногда используется при определении достаточного числа факторов перед вращением. При этом руководствуются следующим правилом: оставлять нужно лишь те факторы, которым соответствуют первые точки на графике до того, как кривая станет более пологой. Отметим, что по умолчанию SPSS

вращает все факторы, чьи собственные значения превышают 1; в данном примере число таких факторов равно 3, а в соответствии с упомянутым правилом достаточно было бы не трех, а четырех факторов.

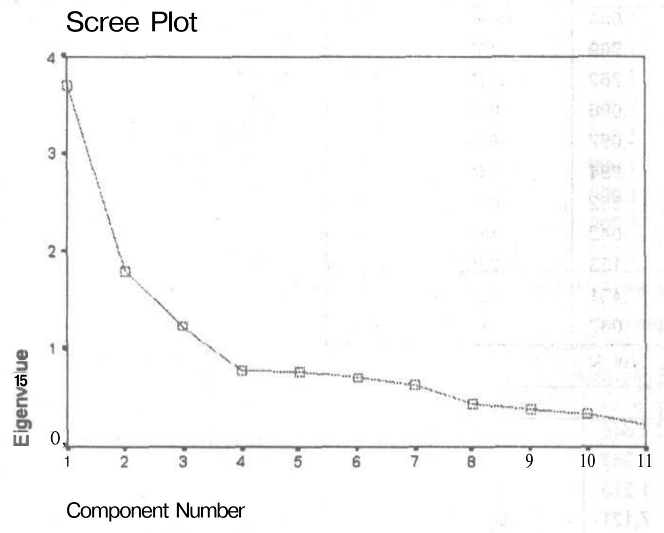



**Рис. 20.10.** График собственных значений

Далее SPSS включает в вывод исходную структуру факторных нагрузок (до вращения). Эти данные в большинстве случаев не представляют интереса, и мы не станем приводить их в этом разделе. На рис. 20.11 показана квадратная матрица преобразований размером 3 × 3.

**Component Transformation Matrix**

| Component | 1 | 2 | 3 |
|---|---|---|---|
| 1 | .713 | .428 | .555 |
| 2 | -.534 | .845 | .034 |
| 3 | -.455 | -.321 | .831 |

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

**Рис. 20.11.** Квадратная матрица преобразований

Если умножить матрицу преобразований на исходную матрицу факторных нагрузок (3 x 11), то в результате получится показанная на рис. 20.12 преобразованная матрица факторных нагрузок.

Обратите внимание, что благодаря установке флажка Sorted by size (Сортировать по величине) в окне Factor Analysis: Options (Факторный анализ: Параметры) факторные нагрузки отсортированы следующим образом:

- наибольшие значения нагрузок для каждого фактора выделяются и сортируются в отдельных блоках;
- внутри каждого блока нагрузки факторов упорядочены по убыванию.

**Rotated Component Matrix**[a]

| | Component | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| счет в уме | .804 | -.056 | .195 |
| аналогии | .769 | .183 | .074 |
| числовые ряды | .762 | .217 | -.005 |
| умозаключения | .696 | .007 | .284 |
| заучивание слов | -.057 | .829 | .048 |
| осведомленность | .094 | .746 | .077 |
| пропущенные слова | .352 | .630 | .202 |
| понятливость | .042 | .155 | .739 |
| скрытые фигуры | .153 | .225 | .737 |
| геометрическое сложение | .421 | -.224 | .692 |
| исключение изображений | .087 | .487 | .506 |

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged In 5 Iterations.

**Рис.** 20.12. Преобразованная матрица факторных нагрузок

Если бы описанный анализ провел реальный исследователь, он был бы удовлетворен полученными результатами. Как нетрудно видеть, первый из факторов соответствует предполагаемым математическим способностям, так как объединяет субтесты «счет в уме», «аналогии», «числовые ряды» и «умозаключения».

Во второй фактор попали три субтеста, относящиеся к вербальным способностям: «заучивание слов», «осведомленность», «пропущенные слова», а в третий фактор — три субтеста, относящиеся к невербальным способностям: «скрытые фигуры», «геометрическое сложение», «исключение изображений». К «странностям» результатов можно отнести разве что распределение переменной «исключение изображений» между вторым и третьим фактором и попадание переменной «понятливость» в третий фактор. Подобные отклонения обычно требуют отдельного изучения. В частности, можно увеличить число факторов или исключить «неопределенные» переменные и повторить анализ.

Целью приведенного примера было показать, каким образом факторный анализ группирует переменные, объединяя их по факторам. Каждый фактор интерпретируется как причина совместной изменчивости (корреляции) группы переменных. Отметим напоследок, что чаще всего исследователь не ограничивается однократной факторизацией данных, а получает несколько вариантов решения с разными наборами переменных и разным числом факторов. Затем выбирается то решение, которое является наилучшим по признакам простоты структуры и концептуальной осмысленности.

# 21 Многомерное шкалирование



Основное достоинство многомерного шкалирования — представление больших массивов данных о различии объектов в наглядном, доступном для интерпретации графическом виде. При многомерном шкалировании матрица различий между объектами (вычисленными, например, по их экспертным оценкам) представляется в виде одно-, двух- или трехмерного графического изображения взаимного расположения этих объектов. Хотя доступно и более трех измерений, эта возможность редко применяется на практике.

Основным преимуществом многомерного шкалирования является возможность очень наглядного визуального сравнения объектов анализа. Если две точки на изображении удалены друг от друга, то между соответствующими объектами имеется значительное расхождение; напротив, близость точек говорит о сходстве объектов.

Многомерное шкалирование имеет много общих черт с факторным анализом (см. главу 20). Так же как и при факторном анализе, создается система координат пространства, в котором определяется расположение точек. Так же как и при факторном анализе, происходит снижение размерности и упрощение данных. Однако при факторном анализе обычно используются коэффициенты корреляции, а при многомерном шкалировании — меры различия между объектами. Наконец, в факторном анализе наибольший интерес вызывают углы между точками, представляющими данные, а в многомерном шкалировании ключевой величиной является расстояние между этими точками.

Помимо факторного анализа многомерное шкалирование имеет несколько общих черт с кластерным анализом (см. главу 22). В обоих случаях анализируется расстояние между объектами; однако при кластерном анализе типичной являет-

ся количественная процедура объединения объектов в группы (кластеры), а при многомерном шкалировании качественный анализ объектов проводится визуально с помощью диаграммы.

Процедура многомерного шкалирования SPSS, имеющая историческое название ALSCAL, фактически не является одной программой, а представляет собой набор небольших процедур, каждая из которых соответствует своему типу данных. В этой главе мы проведем несколько анализов для различных типов данных, стараясь, но возможности, коротко и ясно изложить смысл сопутствующих терминов. Если вы сталкиваетесь с многомерным шкалированием впервые, то рекомендуем обратиться к дополнительной литературе.

В первом примере мы обработаем гипотетическую социограмму для группы учащихся. В этом примере их количественные оценки отношений друг к другу будут преобразованы в графическое изображение взаимного расположения учащихся. Во втором примере мы рассмотрим результаты тестирования учащихся по пяти показателям и графически представим различия между учащимися на плоском изображении. Наконец, третий пример будет представлять собой небольшое исследование восприятия и понимания студентами пяти многомерных методов статистического анализа.

# Квадратная асимметричная матрица различий

Представим себе, что преподаватель решил создать идеальную психологическую обстановку в группе во время занятия, рассадив учащихся так, чтобы ни один из них не оказался рядом с тем, кто ему не нравится. Для этого каждому из 12 студентов было предложено оценить степень своей симпатии к своим однокурсникам по пятибалльной шкале (от 1 до 5, где 1 — максимум симпатии, а 5 — максимум антипатии). Результаты этого вымышленного опроса мы поместили в файл данных mds1.sav. Чтобы добиться желаемого результата, преподавателю необходимо максимально далеко рассадить негативно настроенных в отношении друг друга учащихся. Здесь весьма полезной окажется диаграмма, на которой удаленность точек будет соответствовать отношениям между учащимися. Для построения диаграммы мы воспользуемся средствами многомерного шкалирования.

Первое, что необходимо сделать для решения задачи, — создать квадратную (12 x 12) матрицу различий. Позже на основе этой матрицы будет построено двухмерное изображение, иллюстрирующее взаимоотношения студентов. В ходе многомерного шкалирования исходная матрица 12 x 12 преобразуется в гораздо более простую матрицу 12 x 2 (где 2 — число измерений), содержащую координаты точек для изображения. Исходную матрицу называют *квадратной*

асимметричной матрицей различий. Поясним, что означают составляющие это определение термины.

- *Квадратная матрица* — это матрица, строки и столбцы которой представляют один и тот же набор объектов. В данном случае этим набором объектов является группа учащихся: каждый учащийся оценивает меру своей симпатии-антипатии к каждому из остальных. Если бы строки и столбцы матрицы представляли не одни и те же, а разные объекты, матрица бы называлась не квадратной, а прямоугольной, которая в контексте данной книги не рассматривается.
- *Асимметричная матрица* — это матрица, для которой отношение двух объектов друг к другу может быть разным. Так, например, симпатия Петра к Ирине не означает, что Ирине Петр тоже симпатичен. Визуально асимметричность матрицы выражается в том, что как минимум для одной пары ячеек, симметрично расположенных относительно главной диагонали матрицы, значения различны. Если же ни одного такого различия для матрицы не установлено, то матрица называется симметричной. Примером матрицы, которая всегда симметрична, является корреляционная матрица.
- *Матрица различий* — матрица, данные которой представляют меру различия. В данном случае значения матрицы отражают степень отличия отношения одного студента к другому от идеального; чем больше значение, тем больше различие. Существуют также матрицы сходств, в которых значения отражают степень некоторого сходства. Если бы преподаватель вместо матрицы различий составил матрицу сходств, то изображение, полученное в результате многомерного шкалирования, скорее внесло бы путаницу, чем помогло в решении задачи.

# Квадратная симметричная матрица различий

Пусть наш гипотетический преподаватель решает другую задачу: ему нужно рассадить 12 учащихся в соответствии с результатами их тестирования по пяти показателям. Поскольку результаты тестирования не относятся к данным, характеризующим различия, необходимо сначала вычислить различия по имеющимся данным и таким образом свести задачу к предыдущей. Исходные данные для этой задачи естественно представить в виде прямоугольной матрицы 12 х 5, в которой для каждого из 12 учащихся указаны результаты 5 тестов (файл mds2.sav). Затем по исходным данным строится квадратная (12 × 12) матрица различий между учащимися. Наконец, как и в предыдущем примере, SPSS создает матрицу координат 12 х 2 и визуально представляет ее в виде диаграммы.

Обратите внимание на два ключевых свойства матрицы различий: она является *квадратной* и *симметричной*. Несмотря на то что исходная матрица является прямоугольной, то есть ее строки (объекты) соответствуют учащимся, а столбцы (переменные) — тестам, в матрице различий как строки, так и столбцы соответ-

ствуют учащимся, и, следовательно, матрица является квадратной с размером 12 x 12. Далее, поскольку, к примеру, учащийся 1 отличается от учащегося 5 но результатам тестирования так же, как учащийся 5 от учащегося 1, матрица различий является симметричной.

# Модель индивидуальных различий

В обоих рассмотренных примерах создавалась единственная матрица различий, на основе которой вычислялись координаты точек-объектов для визуального представления. Однако в некоторых случаях возможно координатное представление нескольких матриц различий; примером может служить ситуация, когда несколько экспертов оценивают один и тот же набор объектов. В этих случаях говорят о многомерном шкалировании индивидуальных различий.

Предположим, преподаватель предлагает студентам сравнить между собой 5 многомерных статистических методов: множественный регрессионный анализ (mra), факторный анализ (fa), кластерный анализ (ka), дискриминантный анализ (da) и многомерное шкалирование (mds). Каждому студенту предъявляются все возможные пары из этих методов для оценки различия между ними (1 — максимально сходны ... 5 — максимально различны). Затем по данным для 6 студентов составляется 6 матриц различий ($5 \times 5$ каждая).

Поскольку SPSS предъявляет определенные требования к формату данных при использовании модели индивидуальных различий, мы создали специальный файл mds3.sav. Требования к формату файла, по сути, сводятся к наличию в этом файле нескольких квадратных матриц одинакового размера. Поскольку в данном примере мы имеем дело с 6 студентами, оценивающими 5 объектов (многомерных методов), в файле mds3.sav присутствует 6 матриц размером $5 \times 5$. Первые 5 строк файла соответствуют первой матрице, следующие 5 строк — второй матрице, и т. д. Всего под матрицы отведено 30 строк.

# Пошаговые алгоритмы вычислений

Для проведения многомерного шкалирования сначала необходимо выполнить три подготовительных шага.

**Шаг 1** Создайте новый файл данных или подготовьте существующий. Все необходимые действия описаны в главе 3.

**Шаг 2** Запустите программу SPSS при помощи значка на рабочем столе или команды Пуск ▸ Программы ▸ SPSS for Windows ▸ SPSS 11.5 for Windows (Start ▸ Programs ▸ SPSS for Windows ▸ SPSS 11.5 for Windows) главного меню Windows. В открывшемся после запуска программы диалоговом окне SPSS for Windows щелкните на кнопке Cancel (Отмена).

После выполнения этого шага на экране появится окно редактора данных SPSS.

Шаг 3 Откройте файл данных, с которым вы намерены работать (в нашем случае это файл mds1 .sav). Если он расположен в текущей папке, то выполните следующие действия:

1. Выберите в меню File (Файл) команду Open > Data (Открытие ▸ Данные) или щелкните на кнопке Open File (Открытие файла) панели инструментов.
2. В открывшемся диалоговом окне дважды щелкните на имени mds1.sav или введите его с клавиатуры и щелкните па кнопке OK.

Независимо от того, открыта программа SPSS только что или какие-то процедуры уже выполнялись, в верхней части главного окна должна присутствовать строка меню (она показана ниже). Пока строка меню присутствует на экране, доступны все команды анализа данных. При этом окно с данными видеть не обязательно.

File Edit View Data Transform Analyze Graphs Utilities Window Help

При работе со сводными таблицами или при редактировании диаграмм некоторые пункты строки меню могут исчезать или меняться. Чтобы вернуться к главному окну и стандартной строке меню, щелкните на кнопке свертывания или восстановления текущего окна.

После завершения шага 3 на экране должно присутствовать окно редактора данных.

Шаг 4 В меню Analyze (Анализ) выберите команду Scale ▸ Multidimensional Scaling (Шкалирование ▸ Многомерное шкалирование). Откроется диалоговое окно Multidimensional Scaling (Многомерное шкалирование), показанное на рис. 21.1.

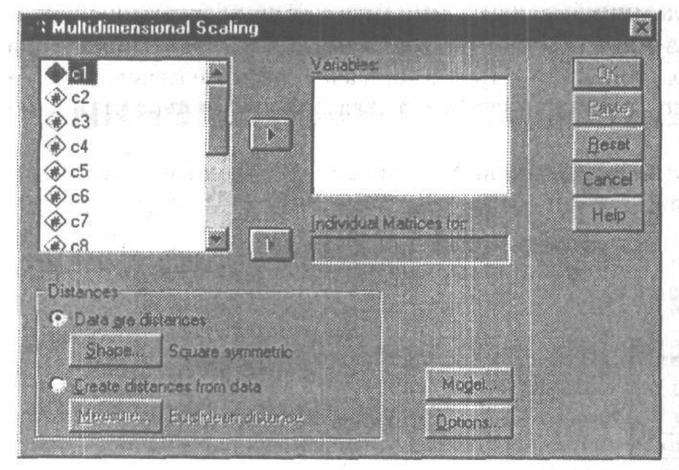



**Рис. 21.1.** Диалоговое окно Multidimensional Scaling

В диалоговом окне Multidimensional Scaling (Многомерное шкалирование) вы можете задать список переменных, которые будут участвовать в анализе. В SPSS существует два типа многомерного шкалирования: с заданной матрицей различий и с вычислением матрицы различий.

В том случае, когда матрица различий задана, вы при желании можете указать столбцы файла, которые будут участвовать в анализе. Обратите внимание на следующий нюанс: вы можете включить в процедуру многомерного шкалирования часть столбцов файла данных, однако не можете включить часть строк. Последнее достигается косвенно — при помощи описанной подробно в главе 4 команды Select Cases (Выбор объектов), однако вы не можете исключить «лишние» объекты непосредственно. Если используется квадратная матрица, то число выбранных переменных (столбцов) должно совпадать с числом объектов файла (строк).

В случае если вы не задаете матрицу различий в файле данных и хотите, чтобы программа SPSS создала ее самостоятельно, требуется указать переменные, для которых будут вычислены расстояния между объектами. В обоих случаях выбор переменных сводится к одной и той же процедуре: вы выделяете нужную переменную (или несколько последовательных переменных) в исходном списке в левой части диалогового окна и щелчком на верхней кнопке со стрелкой помещаете ее в список Variables (Переменные).

Если вы хотите, чтобы программа SPSS вычислила расстояния между объектами вашего файла, и одновременно намереваетесь использовать модель многомерного шкалирования для нескольких выборок, то можете задать имя нужной переменной в поле Individual Matrices for (Отдельные матрицы для), однако эта процедура весьма трудоемка, и мы не будем на ней останавливаться.

После того как заданы переменные, которые должны быть включены в анализ, следует информировать программу SPSS о структуре вашего файла данных. Для этого предназначена группа переключателей Distances (Расстояния). Если ваш файл уже содержит матрицу различий, следует, оставить установленным переключатель Data are distances (Данные являются расстояниями). В случае если матрица с данными является квадратной и симметричной (значения над ее главной диагональю равны значениям под главной диагональю), вам не нужно дополнительно определять тип данных. Если же вы намерены использовать асимметричную или прямоугольную матрицу, щелкните на кнопке Shape (Форма), и на экране появится диалоговое окно Multidimensional Scaling: Shape Of Data (Многомерное шкалирование: Форма данных), представленное на рис. 21.2.

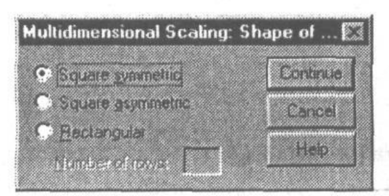



**Рис. 21.2.** Диалоговое окно Multidimensional Scaling: Shape Of Data

В этом окне вы можете выбрать один из возможных типов матрицы. Если ваша матрица является квадратной (число строк равно числу столбцов) и симметричной (значения над главной диагональю равны значениям под главной диагональю), то следует оставить установленным переключатель Square symmetric (Квадратная симметричная). Этот тип матрицы используется в примере с многомерными методами, поскольку сравнения одного метода с другим и наоборот дают одинаковый результат. Если матрица является квадратной, но не симметричной, следует установить переключатель Square asymmetric (Квадратная асимметричная). Такой тип матрицы используется в примере с социограммой студентов. Наконец, если ваша матрица не является квадратной, установите переключатель Rectangular (Прямоугольная) и введите число строк в расположенное рядом поле. Этот тип матрицы используется реже, например при шкалировании предпочтений, и в данной книге не описывается.

В случае если необходимо, чтобы программа SPSS создала матрицу различий по имеющимся данным, в окне Multidimensional Scaling (Многомерное шкалирование) следует установить переключатель Create distances from data (Вычислить расстояния по данным). В ответ SPSS создаст матрицу различий для тех переменных, которые вы указали в списке Variables (Переменные). По умолчанию вычисляется Евклидово расстояние (по формуле Пифагора $a^2 + b^2 = c^2$), и если такой способ вычисления устраивает вас, то никаких дополнительных действий с вашей стороны не требуется. Если же вы хотите применить иной способ вычисления расстояния, то следует щелкнуть па кнопке Measure (Мера), открыв диалоговое окно Multidimensional Scaling: Create Measure from Data (Многомерное шкалирование: Создание меры для данных), представленное на рис. 21.3.

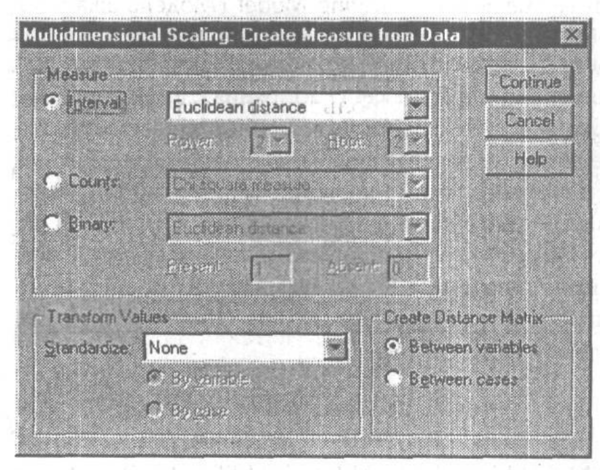



Рис. 21.3. Диалоговое окно Multidimensional Scaling: Create Measure from Data

Как можно видеть, диалоговое окно содержит 3 области: Measure (Мера), Transform Values (Трансформация значений) и Create Distance Matrix (Создание матрицы расстояний). Помимо формулы для Евклидова расстояния существует еще несколько

довольно распространенных формул, например формула Минковского, которая напоминает формулу Пифагора, однако позволяет вместо квадратов использовать любую степень, допустим, четвертую: $a^4 + b^4 = c^4$. Кроме этого, полезным способом определения расстояния является метрика города. Для того чтобы понять ее суть, представьте себе городской район, в котором вы можете передвигаться только по улицам, огибая дома. Метрика города применяется при использовании разнородных по смыслу переменных.

Если все переменные вашего файла имеют одну и ту же шкалу (например, целые числа от 1 до *5),* то, вероятно, вам не придется вместо пункта None (Нет), выбранного в списке Standardize (Стандартизация) но умолчанию, выбирать что-то другое. Если же для переменных применяются различные шкалы, то вам может понадобиться какой-либо из вариантов стандартизации, предлагаемых SPSS. Чаще всего используется вариант Z scores (*z*-шкала), хотя существуют и другие варианты стандартизации значений. Наконец, по умолчанию SPSS вычисляет расстояния между переменными, однако при желании вы можете потребовать вычислить расстояние между объектами. Для этого в группе Create Distance Matrix (Создание матрицы расстояний) необходимо вместо переключателя Between variables (Между переменными) установить переключатель Between cases (Междуобъектами).

После задания режима вычисления расстояний вы можете выбрать желаемый тип модели многомерного шкалирования. Для этого в окне Multidimensional Scaling (Многомерное шкалирование) щелкните на кнопке Model (Модель), и на экране появится диалоговое окно Multidimensional Scaling: Model (Многомерное шкалирование: Модель), представленное на рис. 21.4. Наибольший интерес в этом окне представляют области Dimensions (Измерения) и Scaling Model (Модель шкалирования). В отличие от факторного анализа (см. главу 20), при многомерном шкалировании программа не может автоматически определить количество шкал, которое следует использовать. Поскольку в подавляющем большинстве случаев требуется не более трех шкал, выбор нужного числа измерений не представляет трудностей: вы легко можете перебрать все варианты представления и остановиться на том, которые наиболее удобен для зрительного восприятия. Чаще всего при многомерном шкалировании применяют модель Евклидовых расстояний; в ней вам необходимо задать либо одну матрицу различий, либо несколько идентичных матриц для анализа. Однако если матрицы различий получены от разных субъектов, которых вы хотите сравнить между собой, вам придется использовать модель индивидуальных различий. Примером служит ситуация с упоминавшимися ранее многомерными методами: каждый из шести студентов сравнивает 5 методов, оценивая различия между ними. Группа переключателей Conditionality (Условие) позволяет выбрать, какие условия сравнения следует выполнять: если установлен переключатель Matrix (Матрица), то ячейки каждой матрицы сравниваются между собой (эта ситуация соответствует примеру с методами, когда сравниваются различия, оцененные одним студентом, но не сравниваются различия из разных таблиц). Установка переключателя Row (Строка) означает, что сравнения

производятся между ячейками одной строки. Этот переключатель следует установить в примере с социограммой, поскольку в нем каждая строка таблицы различий соответствует субъекту, оценивающему свое отношение к однокурсникам; в общем случае учащиеся могут использовать разные шкалы, однако сравнения в рамках одной строки имеют смысл. Наконец, переключатель Unconditional (Любые) означает, что в программе допускаются сравнения любых ячеек.

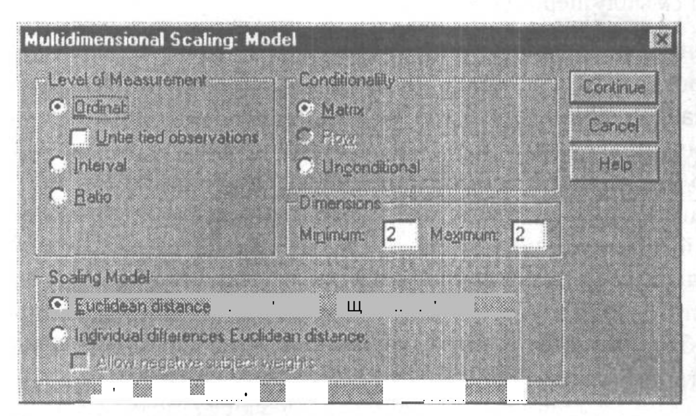



**Рис. 21.4.** Диалоговое окно Multidimensional Scaling: Model

Группа переключателей Level of Measurement (Уровень измерения) позволяет указать программе, в какой шкале интерпретировать величины матрицы различий. Как правило, между переключателями Ordinal (Порядковый), Interval (Интервальный) и Ratio (Отношений) нет существенных различий.

Последним рассматриваемым элементом интерфейса диалогового окна Multidimensional Scaling (Многомерное шкалирование) является кнопка Options (Параметры). При щелчке на ней открывается диалоговое окно Multidimensional Scaling: Options (Многомерное шкалирование: Параметры), представленное на рис. 21.5. Наибольший интерес здесь представляет группа флажков Display (Отображать). Флажок Group plots (Групповые диаграммы) практически всегда следует устанавливать, поскольку в противном случае вам придется рисовать изображение вручную. Флажок Data Matrix (Матрица данных) позволяет выводить па экран исходные данные; эта возможность оказывается весьма полезной в случае, если вы используете SPSS для создания матрицы различий. Флажок Model and options summary (Модель и параметры) позволяет включить в выводимые данные тип и параметры выбранной модели. Остальные элементы интерфейса диалогового окна Multidimensional Scaling: Options (Многомерное шкалирование: Параметры) используются очень редко.

Следующий пример иллюстрирует практическую реализацию двухмерного шкалирования для заданной матрицы различий. Подробное описание задачи приведено в разделе «Квадратная асимметричная матрица различий».

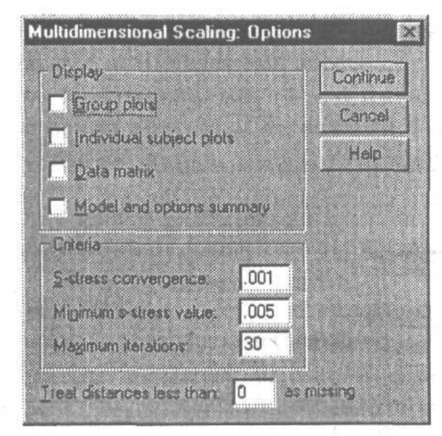



**Рис. 21.5.** Диалоговое окно Multidimensional Scaling: Options

**Шаг 5** После выполнения предыдущего шага у вас должно быть открыто диалоговое окно Multidimensional Scaling (Многомерное шкалирование), представленное на рис. 21.1. Если вы уже успели поработать с этим окном, щелкните на кнопке Reset (Сброс).

1. Щелкпите иа переменной c1, нажмите клавишу Shift и, не отпуская ее, щелкните на перемсппой c12. В результате окажутся выделенными все промежуточные переменные, начиная от переменной c1 и заканчивая переменной c12.
2. Щелкпите на верхней кнопке со стрелкой, чтобы переместить выделенные переменные в список Variables (Переменные).
3. Щелкните на кнопке Shape (Форма), чтобы открыть диалоговое окно Multidimensional Scaling: Shape Of Data (Многомерное шкалирование: Форма данных), показанное на рис. 21.2.
4. Установите переключатель Square asymmetric (Квадратная асимметричная) и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Multidimensional Scaling (Многомерное шкалирование).
5. Щелкните на кнопке Model (Модель), чтобы открыть диалоговое окно Multidimensional Scaling: Model (Многомерное шкалирование: Модель), показанное на рис. 21.4.
6. В группе Conditionality (Условие) установите переключатель Row (Строка) и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Multidimensional Scaling (Многомерное шкалирование).
7. Щелкните иа кнопке Options (Параметры), чтобы открыть диалоговое окно Multidimensional Scaling: Options (Многомерное шкалирование: Параметры), показанное иа рис. 21.5.
8. Установите флажок Group plots (Групповые диаграммы) и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Multidimensional Scaling (Многомерное шкалирование).
9. Щелкните на кнопке OK, чтобы открыть окно вывода.

В следующем примере демонстрируется двухмерное шкалирование квадратной симметричной матрицы различий, которую SPSS создает при задании переменных из файла данных. Данные матрицы различий имеют интервальный тип. Подробное описание задачи было приведено в разделе «Квадратная симметричная матрица различий». В этом примере используется файл данных mds2.sav.

**Шаг 3а** Откройте файл данных, с которым вы намерены работать (в пашем случае это файл mds2.sav). Если он расположен в текущей папке, то выполните следующие действия:

1. Выберите в меню File (Файл) команду Open ▸ Data (Открытие ▸ Данные) или щелкните па кнопке Open File (Открытие файла) панели инструментов.
2. В открывшемся диалоговом окне дважды щелкните на имени mds2.sav или введите сто с клавиатуры и щелкните на кнопке OK.

**Шаг 4а** В меню Analyze (Анализ) выберите команду Scale ▸ Multidimensional Scaling (Шкалирование ▸ Многомерное шкалирование). Откроется диалоговое окно Multidimensional Scaling (Многомерное шкалирование), показанное на рис. 21.1.

**Шаг 5а** После выполнения шага 4а у вас должно быть открыто диалоговое окно Multidimensional Scaling (Многомерное шкалирование). Если вы уже успели поработать с этим окном, щелкните на кнопке Reset (Сброс).

1. Щелкните па переменной тест1, нажмите клавишу Shift и, не отпуская ее, щелкните на переменной тест5. В результате окажутся выделенными все промежуточные переменные, начиная от переменной тест1 и закапчивая переменной тест5.
2. Щелкните на верхней кнопке со стрелкой, чтобы переместить выделенные переменные в список Variables (Переменные).
3. В группе Distances (Расстояния) установите переключатель Create distances from data (Вычислить расстояния по данным) и щелкните на кнопке Measure (Мера), чтобы открыть диалоговое окно Multidimensional Scaling: Create Measure from Data (Многомерное шкалирование: Создание меры для данных), показанное па рис. 21.3.
4. В группе Create Distance Matrix (Создание матрицы расстояний) установите переключатель Between cases (Между объектами) и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Multidimensional Scaling (Многомерное шкалирование).
5. Щелкните на кнопке Model (Модель), чтобы открыть диалоговое окно Multidimensional Scaling: Model (Многомерное шкалирование: Модель), показанное на рис. 21.4.
6. В группе Level of Measurement (Уровень измерения) установите переключатель Interval (Интервальный) и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Multidimensional Scaling (Многомерное шкалирование).
7. Щелкните на кнопке Options (Параметры), чтобы открыть диалоговое окно Multidimensional Scaling: Options (Многомерное шкалирование: Параметры), показанное на рис. 21.5.

8. Установите флажок Group plots (Групповые диаграммы) и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Multidimensional Scaling (Многомерное шкалирование).
9. Щелкните на кнопке OK, чтобы открыть окно вывода.

Последним мы рассмотрим пример двухмерного шкалирования с использованием квадратных симметричных матриц и модели индивидуальных различий. Подробное описание было приведено в разделе «Модель индивидуальных различий» этой главы. Обратите внимание, что для выполнения анализа используется файл mds3.sav.

**Шаг 3б** Откройте файл данных, с которым вы намерены работать (в нашем случае это файл mds3.sav). Если он расположен в текущей папке, то выполните следующие действия:

1. Выберите в меню File (Файл) команду Open ▸ Data (Открытие > Данные) или щелкните на кнопке Open File (Открытие файла) панели инструментов.
2. В открывшемся диалоговом окне дважды щелкните на имени mds3.sav или введите его с клавиатуры и щелкните на кнопке OK.

**Шаг 4б** В меню Analyze (Анализ) выберите команду Scale ▸ Multidimensional Scaling (Шкалирование ▸ Многомерное шкалирование). Откроется диалоговое окно Multidimensional Scaling (Многомерное шкалирование), показанное на рис. 21.1.

**Шаг 5б** После выполнения предыдущего шага у вас должно быть открыто диалоговое окно Multidimensional Scaling (Многомерное шкалирование). Если вы уже успели поработать с этим окном, щелкните на кнопке Reset (Сброс).

1. Щелкните на переменной mra, нажмите клавишу Shift и, не отпуская ее, щелкните на переменной mds. В результате окажутся выделенными все промежуточные переменные, начиная от переменной mra и заканчивая переменной mds.
2. Щелкните на верхней кнопке со стрелкой, чтобы переместить выделенные переменные в список Variables (Переменные).
3. Щелкните на кнопке Model (Модель), чтобы открыть диалоговое окно Multidimensional Scaling: Model (Многомерное шкалирование: Модель), показанное на рис. 21.4.
4. В группе Level of Measurement (Уровень измерения) установите переключатель Ratio (Отношение), а в группе Scaling Model (Модель шкалирования) — переключатель Individual differences Euclidian distance (Индивидуальные различия. Евклидово расстояние), после чего щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Multidimensional Scaling (Многомерное шкалирование).
5. Щелкните на кнопке Options (Параметры), чтобы открыть диалоговое окно Multidimensional Scaling: Options (Многомерное шкалирование: Параметры), показанное на рис. 21.5.
6. Установите флажок Group plots (Групповые диаграммы) и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Multidimensional Scaling (Многомерное шкалирование).
7. Щелкните на кнопке OK, чтобы открыть окно вывода.

После выполнения шага 5 программа автоматически активизирует окно вывода. Для просмотра результатов вы при необходимости можете воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

# Печать результатов и выход из программы

Ниже описана типичная процедура печати результатов статистического анализа (или нескольких анализов). После выполнения шага 5 должно быть открыто окно вывода.

Шаг 6 В окне вывода укажите фрагменты, выводимые на печать (см. раздел «Окно вывода» в главе 2), в меню File (Файл) выберите команду Print (Печать), при необходимости задайте параметры печати и щелкните на кнопке OK.

Последнее, что необходимо сделать после завершения исследования и печати результатов, — это выйти из программы SPSS.

Шаг 7 Для выхода из программы в меню File (Файл) выберите команду Exit (Выход).

Иногда после выполнения команды Exit (Выход) на экране могут появляться небольшие диалоговые окна с вопросом о необходимости сохранения сделанных в файлах изменений и кнопками, описывающими возможные варианты ответа. Для завершения работы просто щелкайте па соответствующих кнопках.

# Представление результатов

В этом разделе приведены данные, сгенерированные командой Multidimensional Scaling (Многомерное шкалирование) при выполнении шага 5.

## Итерации

Значения, записанные в столбце S-Stress, характеризуют отклонение результата от идеального (точно соответствующего матрице отличий) на различных итерациях применения модели (рис. 21.6). SPSS применяет заданную модель столько раз, сколько необходимо для получения достаточно низкого значения в столбце S-Stress. Если число итераций оказывается больше 30, то это, как правило, указывает на проблемы в исходных данных.

```
Iteration history for the 2 dimensional solution (in squared distances)

                 Young's S-stress formula 1 is used.

               Iteration      S-stress        Improvement

                   1           .36850
                   2           .34885           .01965
                   3           .33645           .01240
                   4           .32766           .00879
                   5           .32256           .00509
                   6           .32076           .00181
                   7           .32003           .00072

                     Iterations stopped because
                 S-stress improvement is less than    .001000
```

Рис. 21.6. Фрагмент окна вывода после выполнения шага 5 (итерации)

## Стрессы и квадраты коэффициентов корреляции

Для каждой строки асимметричной матрицы различий, для каждой матрицы модели индивидуальных различий, а также для всей модели при многомерном шкалировании вычисляются стресс и коэффициент $R^2$ (рис. 21.7). Стресс но своему смыслу схож со стрессом предыдущей модели, однако для его расчета используется другое уравнение, позволяющее упростить вычислительный процесс сравнения различий. Коэффициент $R^2$ (столбец RSQ) характеризует долю дисперсии в матрице различий, обусловленную данной моделью. Чем лучше модель, тем выше значение коэффициента $R^2$. Если вы, к примеру, строите несколько графических изображений с разным числом шкал-координат, то величины стресса и $R^2$ могут служить критериями при выборе наиболее подходящей модели.

```
      Stress and squared correlation (RSQ) in distances
                          Matrix     1
   Stimulus     Stress       RSQ   Stimulus     Stress      RSQ

      1          .225       .826      2          .293      .690
      3          .222       .846      4          .360      .380
      5          .239       .750      6          .188      .839
      7          .206       .796      8          .256      .713
      9          .249       .716     10          .344      .607
     11          .250       .722     12          .162      .880

    Averaged (rms) over stimuli
 Stress  •   .256     RSQ =  .730
```

Рис. 21.7. Фрагмент окна вывода после выполнения шага 5 (стрессы и квадраты коэффициентов корреляции)

## Координаты стимулов

Для каждого шкалируемого объекта указываются его координаты по каждой шкале (рис. 21.8). Это сделано для того, чтобы вы могли на основе этих координат

построить собственное графическое изображение или использовать координаты для дальнейшего анализа. В данном случае левый столбец соответствует координате *x*, а правый — координате *y*.

Stimulus Coordinates

| Stimulus Number | Stimulus Name | Dimension 1 | 2 |
|---|---|---|---|
| 1 | C1 | .9541 | 1.1290 |
| 2 | C2 | 1.3036 | .2898 |
| 3 | C3 | 1.1314 | .8383 |
| 4 | C4 | -.7832 | .7595 |
| 5 | C5 | -1.6958 | .2091 |
| 6 | C6 | -1.6310 | -.0369 |
| 7 | C7 | -.3179 | -1.5237 |
| 8 | C8 | .0372 | -1.6126 |
| 9 | C9 | -1.0067 | ,7802 |
| 10 | C10 | .8591 | .7622 |
| 11 | C11 | .8739 | -.0226 |
| 12 | C12 | .2754 | -1.5722 |

**Рис. 21.8.** Фрагмент окна вывода после выполнения шага *5* (координаты стимулов)

## Субъективные веса

Если вы используете модель индивидуальных различий, то SPSS включает в вывод подсчитанные значения весов для каждого субъекта (рис. 21.9). В примере со сравнением многомерных методов каждый из 6 студентов по каждой из двух шкал имеет свой вес, который характеризует, какая доля дисперсии субъективных оценок приходится на соответствующую шкалу. Кроме того, в вывод включен общий вес для каждой из шкал. Величина Weirdness (предсказуемость) характеризует разброс весов для каждого субъекта; для субъекта с номером 3 она имеет максимальное значение, равное 0,51.

Subject Weights

| Subject Number | Weird-ness | Dimension 1 | 2 |
|---|---|---|---|
| 1 | .1841 | .7054 | .6774 |
| 2 | .1869 | .7075 | .6825 |
| 3 | .5110 | .8968 | .2596 |
| 4 | .1161 | .7555 | .6500 |
| 5 | .1431 | .7258 | .6521 |
| 6 | .2415 | .8951 | .4346 |
| Overall importance of each dimension: | | .6169 | .3381 |

**Рис. 21.9.** Фрагмент окна вывода после выполнения шага 5б (субъективные веса)

## Итоговая конфигурация объектов

Диаграмма, показанная на рис. 21.10, представляет собой итог применения модели многомерного шкалирования (в результате выполнения шага 5). Она отображает взаимоотношения 12 студентов таким образом, что чем больше различия между учащимися в исходной матрице, тем дальше они находятся друг от друга на диаграмме. С помощью диаграммы мы можем видеть, что в данной группе выделяются три относительно компактные подгруппы, самая крупная из которых состоит из пяти человек и располагается в правом верхнем углу диаграммы. Отношения внутри каждой из группировок характеризуются симпатией (точки расположены близко), чего не скажешь об отношениях между группами. В данном случае смысл каждой из шкал не имеет значения; главным является взаимное расположение точек.

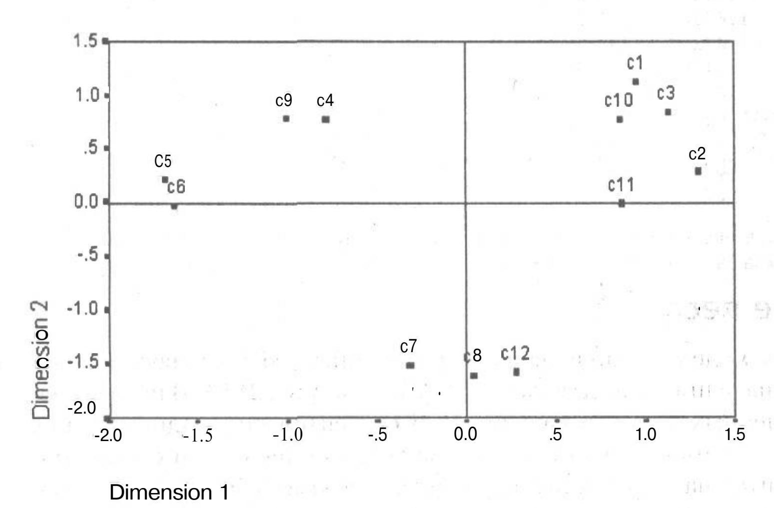



**Рис. 21.10.** Фрагмент окна вывода после выполнения шага 5 (итоговая диаграмма)

На диаграмме, представленной на рис. 21.11, изображена итоговая конфигурация сравниваемых студентами статистических методов (результат выполнения шага 5б). В этом случае можно попытаться дать содержательную интерпретацию выделенных шкал как критериев, которыми пользуются студенты при сравнении методов.

На правом полюсе шкалы 1 расположились множественный регрессионный (mra) и факторный анализы (fa), а на противоположном полюсе этой шкалы — кластерный анализ (ka) и многомерное шкалирование (mds). Эта шкала, таким образом, различает методы по предмету анализа: правый полюс шкалы соответствует корреляциям (факторный и регрессионный анализ), левый — различиям и расстояниям (шкалирование и кластерный анализ). На положительном полюсе шкалы 2 оказались дискриминантный (da) и множественный регрессионный анализы (mra), а на отрицательном полюсе этой шкалы — многомерное

шкалирование (mds) и факторный анализ (fa). По всей видимости, эта шкала разделяет методы по их назначению: положительный полюс соответствует методам предсказания, а отрицательный — структурным методам. О значении этих шкал как субъективных критериев, которыми пользовались студенты при сравнении методов, можно судить по субъективным весам (см. рис. 21.9).

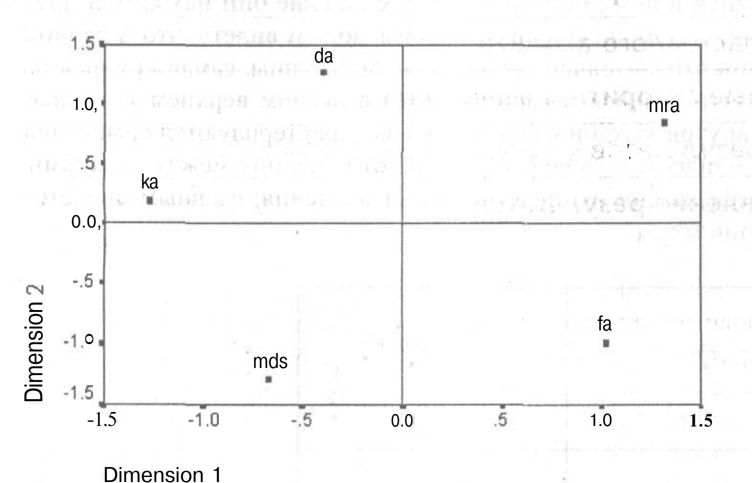



Рис. 21.11. Фрагмент окна вывода после выполнения шага 5б (итоговая диаграмма)

# 22 Кластерный анализ



Зачастую описание нового статистического метода удобно проводить путем его сравнения с другим методом. Именно таким образом мы и начнем рассмотрение иерархического кластерного анализа, сравнивая его с факторным анализом. При многочисленных общих чертах между указанными статистическими методами существует немало различий. Если вы еще не знакомы с понятием факторного анализа, настоятельно рекомендуем вам обратиться к началу главы 20, поскольку далее мы займемся сравнительной характеристикой факторного и кластерного анализов. Как обычно, после теоретической части последуют примеры практической реализации статистических методов средствами SPSS, оформленные в виде пошаговых процедур.

## Сравнение кластерного и факторного анализов

Главное сходство между кластерным и факторным анализами заключается в том, что тот и другой предназначены для перехода от исходной совокупности множества переменных (или объектов) к существенно меньшему числу факторов (кластеров). Тем не менее реализация статистических процедур и интерпретация результатов для двух типов анализа различаются; пять основных отличий приведены далее.

- Целью факторного анализа является замена большого числа исходных переменных меньшим числом факторов. Кластерный анализ, как правило, применяется для того, чтобы уменьшить число объектов путем их группировки. Другими словами, в процедуре кластерного анализа обычно переменные не группируются, а выступают в качестве критериев для группировки объектов. Так, в примере факторного анализа (см. главу 20) 11 субтестов интеллекта (переменных) были сведены к трем факторам, каждый из которых объединил

несколько родственных исходных переменных. Кластерный анализ применяется обычно для выделения групп объектов, исходя из их сходства по измеренным признакам. Применительно к примеру с 11 субтестами интеллекта типичной задачей кластерного анализа была бы классификация учащихся (объектов) таким образом, чтобы по измеренным 11 показателям внутри каждой группы объекты были бы более похожи друг на друга, чем на объекты из других групп. Группы объектов, выделенные в результате кластерного анализа на основе заданной меры сходства между объектами, называются кластерами.

- Заявленные в предыдущем пункте различия между кластерным и факторным вариантами анализа со всей полнотой категоричности могут быть отнесены лишь к ранним версиям SPSS. Начиная с версии SPSS 10.0, программа позволяет с равным успехом проводить кластерный анализ не только объектов, но и переменных. В последнем случае кластерный анализ может выступать как более простой и нередко более эффективный аналог факторного анализа. В разделе пошаговых процедур мы продемонстрируем оба варианта кластерного анализа.
- Действия, выполняемые в ходе статистических операций в каждом из вариантов анализа, принципиально различаются. В факторном анализе на каждом этапе извлечения фактора для каждой переменной подсчитывается доля дисперсии, которая обусловлена влиянием данного фактора. При кластерном анализе вычисляется расстояние между текущим объектом и всеми остальными объектами, и кластер образует та пара, для которой расстояние оказалось наименьшим. Подобным образом каждый объект либо группируется с другим объектом, либо включается в состав существующего кластера. Процесс кластеризации конечен и продолжается до тех пор, пока все объекты не будут объединены в один кластер. Разумеется, подобный результат в общем случае не имеет смысла, и исследователь должен самостоятельно определить, в какой момент кластеризация должна быть прекращена.
- В контексте кластерного анализа особое место занимает один из его видов, называемый *иерархическим кластерным анализом*. В SPSS он реализуется с помощью команды Hierarchical Cluster (Иерархическая кластеризация). Этот вид кластерного анализа чаще используется в экономике, социологии, политологии, нежели в психологии. Психологи обычно анализируют переменные с целью найти статистические связи между ними; эти связи, как правило, указывают на сходство между теми или иными исследуемыми факторами. Деление выборки на группы в психологических анализах редко представляет интерес; в случаях когда это оказывается необходимым, психологи отдают предпочтение дискриминантному, а не кластерному анализу (см. главу 23).
- Поскольку кластеризация переменных оказывается весьма доступной операцией, было бы интересно сравнить ее результаты с результатами более сложного факторного анализа. Как и в случае факторного анализа, выполнение кластерного анализа и его результаты зависят от ряда параметров: способа вычисления расстояния между объектами, кластеризации индивидуальных объектов и т. д.

# Этапы кластерного анализа

Кластерный анализ выполняется за несколько этапов, приводящих к конечному результату. Сначала мы рассмотрим пример, созданный специально, чтобы показать суть кластерного анализа. Отметим, что кластерный анализ неприменим к файлам данных, использовавшимся ранее, поскольку при их составлении основное внимание было уделено смыслу и связям между переменными, а содержимое объектов (то есть информация, касающаяся субъектов) практически не играло роли. Для демонстрации кластерного анализа нами был подготовлен файл cars.sav, содержащий гипотетические данные о 15 подержанных автомобилях разных марок, выставленных на продажу. Файл имеет структуру, подходящую для наглядной иллюстрации кластерного анализа.

Итак, выделяют несколько этапов кластерного анализа.

1. *Выбор переменных-критериев для кластеризации.* В нашем примере кластеризация будет осуществляться по следующим переменным: цена (стоимость), т_сост (экспертная оценка технического состояния по 10-балльной шкале), возраст (количество лет эксплуатации), пробег (пройденный километраж с начала эксплуатации).

2. *Выбор способа измерения расстояния между объектами, или кластерами* (изначально считается, что каждый объект соответствует одному кластеру). По умолчанию используется квадрат Евклидова расстояния, согласно которому расстояние между объектами равно сумме квадратов разностей между значениями одноименных переменных объектов. Предположим, что марка автомобиля А имеет показатели технического состояния и возраста 5 и 6, а марка В — соответственно 7 и 4. В этом случае расстояние между марками вычисляется следующим образом: $(5 - 7)^2 + (6 - 4)^2 = 8$. При выполнении анализа сумма квадратов разностей вычисляется для всех переменных. Получаемые расстояния используются программой при формировании кластеров. Помимо Евклидова существуют и другие виды расстояний, вычисляемые по другим формулам, однако мы не будем на них останавливаться. При необходимости обратитесь к руководству пользователя SPSS.

   Относительно вычисления расстояния может возникнуть следующий вопрос: будет ли адекватным результат кластерного анализа в том случае, если переменные имеют различные шкалы измерения? Так, все переменные файла car.sav имеют самые разные шкалы. Для решения проблемы шкалирования в SPSS используется стандартизация, в частности, ее простой метод — нормализация переменных, приводящая все переменные к стандартной $z$-шкале (среднее равно 0, стандартное отклонение — 1). Помимо одинаковой шкалы нормализованные переменные также имеют равные веса. В случае если все исходные данные имеют одну и ту же шкалу измерения либо веса переменных по смыслу должны быть разными, стандартизацию переменных проводить не нужно. •

3. *Формирование кластеров.* Существует два основных метода формирования кластеров: *метод слияния* и *метод дробления,* В первом случае исходные кластеры увеличиваются путем объединения до тех пор, пока не будет сформирован единственный кластер, содержащий все данные. Метод дробления основан на обратной операции: сначала все данные объединяются в один кластер, который затем делится на части до тех пор, пока не будет достигнут желаемый результат. По умолчанию программой SPSS используется метод слияния, и мы рассмотрим его в этой главе.

   В методе слияния предусмотрено несколько способов объединения объектов. Способ, применяемый по умолчанию, называется *межгрупповым связыванием,* или *связыванием средних внутри групп.* SPSS вычисляет наименьшее среднее значение расстояния между всеми нарами групп и объединяет две группы, оказавшиеся наиболее близкими. На первом шаге, когда все кластеры представляют собой одиночные объекты, данная операция сводится к обычному попарному сравнению расстояний между объектами. Термин «среднее значение» приобретает смысл лишь па втором этапе, когда сформированы кластеры, содержащие более одного объекта. Так, в пашем примере на начальном этапе имеется 15 кластеров (объектов); сначала в кластер объединяются два объекта с наименьшим расстоянием друг от друга. Затем подсчет расстояний повторяется, и в кластер объединяется еще одна пара переменных. На втором этапе вы получите либо 13 свободных объектов и 1 кластер, объединяющий 2 объекта, либо 11 свободных объектов и 2 кластера по 2 объекта в каждом. В конечном счете все объекты окажутся в одном большом кластере. Существуют и другие методы объединения объектов, однако мы не станем рассматривать их в этой книге. При необходимости обратитесь к руководству пользователя SPSS.

4. *Интерпретация результатов.* Как и в случае факторного анализа, желаемое число кластеров и оценка результатов анализа зависят от целей исследователя. Для рассматриваемого примера нам представляется наиболее предпочтительным число кластеров, равное 3. Как показывает анализ, все марки можно разделить на 3 группы: первая группа имеет высокую стоимость (среднее значение — 15 230), небольшой срок эксплуатации (4 года) и средний пробег (85 400 км). Вторая группа имеет среднюю стоимость, небольшой пробег, наибольший возраст, но хорошее техническое состояние. Третья группа содержит недорогие модели с большим пробегом и невысоким рейтингом технического состояния.

# Пошаговые алгоритмы вычислений

Структура раздела пошаговых процедур принципиально ничем не отличается от аналогичных разделов в других главах книги: шаги 1-4 являются подготовительными для выполнения анализа, шаги 6-7 требуются для печати результатов и завершения работы программы, а несколько вариантов (в данном случае два) шага 5 соответствуют разным способам выполнения статистической операции. Сначала мы продемонстрируем вариант использования кластерного анализа объектов

с включением в вывод нескольких интересующих нас величин. Второй вариант шага 5 проиллюстрирует кластерный анализ переменных. В нем мы вновь обратимся к файлу TestIQ.sav для того, чтобы вы имели возможность сравнить кластерный анализ переменных с факторным анализом, описанным в предыдущей главе.

При проведении кластерного анализа сначала выполняются три подготовительных шага.

**Шаг 1** Создайте новый файл данных или подготовьте существующий. Все необходимые действия описаны в главе 3

**Шаг 2** Запустите программу SPSS при помощи значка на рабочем столе или команды Пуск ▸ Программы ▸ SPSS for Windows ▸ SPSS 11.5 for Windows (Start ▸ Programs ▸ SPSS for Windows ▸ SPSS 11.5 for Windows) главного меню Windows. В открывшемся после запуска программы диалоговом окне SPSS for Windows щелкните на кнопке Cancel (Отмена).

После выполнения этого шага на экране появится окно редактора данных SPSS.

**Шаг 3** Откройте файл данных, с которым вы намерены работать (в нашем случае это файл cars.sav). Если он расположен в текущей папке, то выполните следующие действия:

1. Выберите в меню File (Файл) команду Open ▸ Data (Открытие ▸ Данные) или щелкните на кнопке Open File (Открытие файла) панели инструментов.
2. В открывшемся диалоговом окне дважды щелкните на имени cars.sav или введите его с клавиатуры и щелкните на кнопке OK.

Независимо от того, открыта программа SPSS только что или какие-то процедуры уже выполнялись, в верхней части главного окна должна присутствовать строка меню (она показана ниже). Пока строка меню присутствует на экране, доступны все команды анализа данных. При этом окно с данными видеть не обязательно.

File Edit View Data Transform Analyze Graphs Utilities Window Help

При работе со сводными таблицами или при редактировании диаграмм некоторые пункты строки меню могут исчезать или меняться. Чтобы вернуться к главному окну и стандартной строке меню, щелкните на кнопке свертывания или восстановления текущего окна.

После завершения шага 3 на экране должно присутствовать окно редактора данных.

**Шаг 4** В меню Analyze (Анализ) выберите команду Classify ▸ Hierarchical Cluster (Классификация ▸ Иерархическая кластеризация). Откроется диалоговое окно Hierarchical Cluster Analysis (Иерархический кластерный анализ), показанное на рис. 22.1.

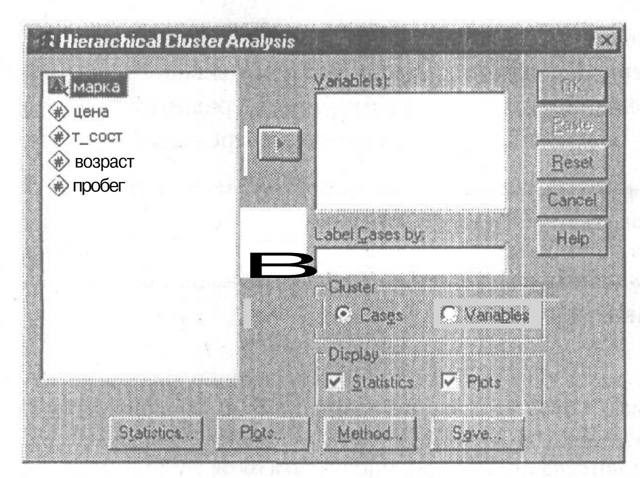



Рис. 22.1. Диалоговое окно Hierarchical Cluster Analysis

Ваши дальнейшие действия в значительной степени зависят от того, какой тип кластеризации вы выберете. Для этой цели в группе Cluster (Кластеризация) предусмотрены два переключателя Cases (Объекты) и Variables (Переменные). В первом случае вам предстоит указать в списке Variable(s) (Переменные) имена тех переменных, значения которых будут использоваться при кластеризации объектов. В нашем примере в список Variable(s) (Переменные) следует поместить все переменные, кроме переменной марка, поскольку последняя представляет собой марку автомобиля. Далее следует задать способ идентификации объектов. Как правило, в роли идентификатора выступает переменная, содержащая уникальный номер объекта или его имя в виде строки. В данном случае мы будем использовать вполне подходящую для этого переменную марка. Имя идентифицирующей переменной указывается в поле Label Cases by (Различать объекты по).

Если вместо переключателя Cases (Объекты) в группе Cluster (Кластеризация) установить переключатель Variables (Переменные), то в списке Variable(s) (Переменные) потребуется указать переменные, помещаемые в кластер, а поле Label Cases by (Различать объекты по) останется пустым. По умолчанию флажки Statistics (Статистики) и Plots (Диаграммы) в группе Display (Отображать) установлены, и в большинстве случаев нет необходимости их сбрасывать. В нижней части диалогового окна расположены четыре кнопки, предназначенные для задания дополнительных параметров команды.

При щелчке на кнопке Statistics (Статистики) на экране появляется диалоговое окно Hierarchical Cluster Analysis: Statistics (Иерархический кластерный анализ: Статистики), представленное на рис. 22.2.

Флажок Agglomeration Schedule (Последовательность слияния) по умолчанию установлен, обеспечивая включение в результаты стандартного компонента вывода кластерного анализа, который будет подробно рассмотрен в разделе «Представление результатов». Флажок Proximity Matrix (Матрица сходства) предназначен

для отображения информации о расстояниях между объектами и кластерами. Использование матрицы удобно лишь для небольших файлов данных, поскольку с увеличением числа объектов размер матрицы резко возрастает, что делает ее громоздкой и неудобной для восприятия. Группа Cluster Membership (Кластеры в решении) состоит из трех переключателей, описанных ниже.

- None (Нет) — в выводимые результаты включаются все кластеры. Этот вариант установлен по умолчанию.
- Single solution (Единственное решение) — позволяет определить точное число выводимых кластеров.
- Range of solutions (Диапазон решений) — обеспечиваетвывод нескольких решений с разным числом кластеров. Так, если ввести в поле From (От) число 3, а в поле through (До) число 5, то в выводимые результаты будут включены все решения с 3, 4 и 5 кластерами.

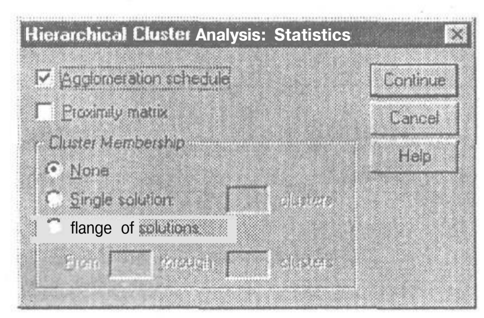



**Рис. 22.2.** Диалоговое окно Hierarchical Cluster Analysis: Statistics

Щелчок на кнопке Plots (Диаграммы) в окне Hierarchical Cluster Analysis (Иерархический кластерный анализ) открывает диалоговое окно Hierarchical Cluster Analysis: Plots (Иерархический кластерный анализ: Диаграммы), представленное на рис. 22.3.

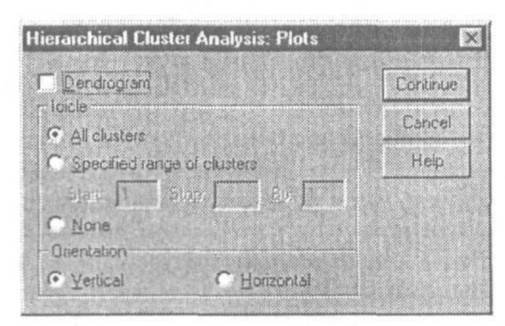



**Рис. 22.3.** Диалоговое окно Hierarchical Cluster Analysis: Plots

Флажок Dendrogram (Дендрограмма) позволяет включить в выводимые результаты ту же информацию, которая содержится на диаграмме накопления, предлагаемой по умолчанию, а также относительную величину разности между переменными

или кластерами на каждом шаге процесса. Поскольку диаграмма накопления (icicle) лишь дублирует информацию других разделов, а ее вид оставляет желать лучшего, мы откажемся от ее вывода. Для этого в группе Icicle (Диаграмма накопления) достаточно установить переключатель None (Нет). Переключатель Vertical (Вертикальная) в группе Orientation (Ориентация) позволяет отобразить на одной странице приблизительно 24 объекта, в то время как переключатель Horizontal (Горизонтальная) - лишь около 14. Последний удобен в том случае, если нужное число объектов, или переменных, размещенных в столбцах, не помещается на одной странице.

Щелчок на кнопке Method (Метод) в окне Hierarchical Cluster Analysis (Иерархический кластерный анализ) открывает диалоговое окно Hierarchical Cluster Analysis: Method (Иерархический кластерный анализ: Метод), представленное на рис. 22.4.

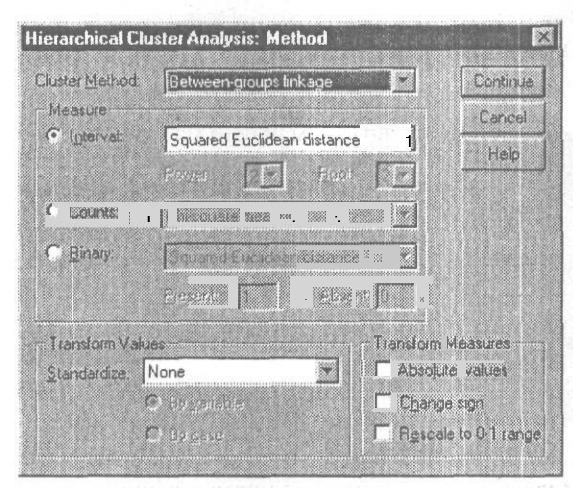



**Рис. 22.4.** Диалоговое окно Hierarchical Cluster Analysis: Method

В этом окне для нас представляют интерес раскрывающиеся списки Cluster Method (Метод кластеризации), Interval (Интервал) и Standardize (Стандартизация). Мы опишем лишь наиболее важные пункты этих списков и кратко упомянем остальные. Более детальные описания вы можете найти самостоятельно в руководстве пользователя SPSS.

В списке Cluster Method (Метод кластеризации) наиболее часто используется пункт Between-groups linkage (Межгрупповое связывание). Другое название этого метода кластеризации — связывание средних внутри групп. Суть метода заключается в том, что объединению на каждом шаге подвергаются кластеры или объекты, расстояние между которыми минимально. Более полное описание межгруппового связывания можно найти в разделе «Представление результатов». Помимо пункта Between-groups linkage (Межгрупповое связывание) раскрывающийся список Cluster Method (Метод кластеризации) содержит следующие пункты:

- Within-groups linkage (Внутригрупповое связывание);
- Nearest neighbor (Одиночное связывание);

- Furthest neighbor (Полное связывание);
- Centroid clustering (Центроидная кластеризация);
- Ward's method (Метод Варда).

В раскрывающемся списке Interval (Интервал) по умолчанию выбран пункт Squared Euclidean distance (Квадрат Евклидова расстояния). Это означает, что расстояние между объектами вычисляется как разность квадратов соответствующих переменных этих объектов, участвующих в анализе. Остальные пункты списка Interval (Интервал):

- Cosine (Косинус) — метод измерения близости, основанный на косинусах векторов значений;
- Pearson correlation (Корреляция Пирсона) — метод измерения близости, основанный на корреляции векторов значений;
- Chebychev (Чебышев) — вычисление расстояния как максимума абсолютной величины разности между элементами;
- Block (Блок) — определяет меру расстояния по метрике города (см. главу 21);
- Minkowsky (Минковский) — определяет меру расстояния Минковского (см. главу 21);
- Customized (Настройка) — позволяет задавать пользовательскую меру расстояния.

Процедура стандартизации, о которой шла речь в начале этой главы, выбирается в раскрывающемся списке Standardize (Стандартизация). По умолчанию выбран пункт None (Нет), однако в случаях, когда стандартизация необходима, чаще всего выбирают пункт Z scores ($z$-шкала). В качестве альтернативы нормализации можно выбрать два следующих по списку пункта (см. далее), однако результат стандартизации во всех трех случаях будет одним и тем же. Оставшиеся два пункта, в которых допускается варьирование среднего значения или стандартного отклонения распределения, могут привести к другим результатам; выбор какого-либо из них определяется степенью его применимости к исследуемым данным и удобством для исследователя. Итак, остальные пункты списка Standardize (Стандартизация):

- Range -1 to 1 (Интервал от –1 до 1) — значения всех переменных заключаются в интервал от -1 до 1;
- Range 0 to 1 (Интервал от 0 до 1) — значения всех переменных заключаются в интервал от 0 до 1;
- Maximum magnitude of 1 (Максимум, равный 1) — распределения изменяются таким образом, что их максимальные значения становятся равными 1;
- Mean of 1 (Среднее, равное 1) — распределения изменяются таким образом, что их средние значения становятся равными 1 (стандартное отклонение может варьировать);

- Standard deviation of 1 (Стандартное отклонение, равное 1) — распределения изменяются таким образом, что их стандартные отклонения становятся равными 1 (среднее значение может варьировать);
- Customized (Настройка) — тип стандартизации определяется пользователем.

В группе Transofm Measures (Преобразование значений) имеется три флажка, позволяющих изменить значения переменных: Absolute values (Абсолютные значения), Change Sign (Смена знаков) и Rescale to 0-1 range (Свести к интервалу 0–1).

Последняя из четырех функциональных кнопок — кнопка Save ( Сохранить) окна Hierarchical Cluster Analysis (Иерархический кластерный анализ) — открывает диалоговое окно Hierarchical Cluster Analysis: Save New Variables (Иерархический кластерный анализ: Сохранение новых переменных), представленное на рис. 22.5. С помощью этого окна можно создавать новые переменные, значения которых будут хранить вычисленные статистические величины. Если установлен переключатель None (Нет), то никакого сохранения в процессе анализа не производится. В противном случае при выполнении анализа будут созданы переменные, которые окажутся в конце файла. Например, если вы установите переключатель Single Solution (Единственное решение) и укажете в поле число 3, то получите новую переменную, значение которой равно 1, 2 или 3 в зависимости от того, какому кластеру будет принадлежать соответствующий объект в решении. Если же установить переключатель Range of solutions (Диапазон решений), в поле From (С) указать число 3, а в поле Through (По) — число 5, то это приведет к созданию трех новых переменных: первая будет принимать значения от 1 до 3, вторая - от 1 до 4, третья — от 1 до 5.

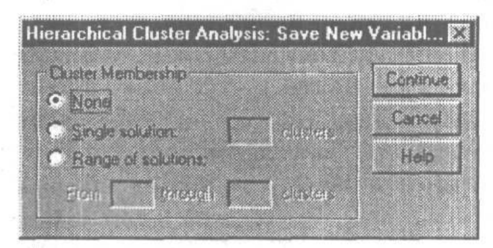



**Рис. 22.5.** Диалоговое окно Hierarchical Cluster Analysis: Save New Variables

В следующем примере проводится кластерный анализ с несколькими дополнительными параметрами, обсуждавшимися выше. В качестве идентификатора используется переменная марка. Все остальные переменные файла задействуются для вычисления расстояния между объектами. Мы включим в выводимые результаты последовательность слияния и дендрограмму, но исключим диаграмму накопления. Значения всех переменных нормализуем для того, чтобы придать им равные веса и привести к одной шкале. В качестве расстояния между объектами зададим квадрат Евклидова расстояния, а в качестве метода кластеризации - межгрупповое связывание. Кроме того, мы создадим новую переменную, в которой сохраним решение с тремя кластерами.

**Шаг 5** После выполнения шага 4 должно быть открыто диалоговое окно Hierarchical Cluster Analysis (Иерархический кластерный анализ), показанное на рис. 22.1.

Если вы уже успели поработать с этим окном, щелкните на кнопке Reset (Сброс).

1. Щелкните сначала на переменной марка, чтобы выделить ее, а затем — на нижней кнопке со стрелкой, чтобы переместить переменную в поле Label Cases by (Различать объекты по).
2. Нажмите кнопку мыши на переменной цена и, не отпуская кнопки, перетащите указатель па переменную пробег, затем кнопку мыши отпустите. В результате окажутся выделенными все оставшиеся в списке переменные.
3. Щелкните на верхней кнопке со стрелкой, чтобы переместить выделенные переменные в список Variable(s) (Переменные).
4. Щелкните на кнопке Plots (Диаграммы), чтобы открыть диалоговое окно Hierarchical Cluster Analysis: Plots (Иерархический кластерный анализ: Диаграммы), показанное на рис. 22.3.
5. Установите флажок Dendrogram (Дендрограмма) и переключатель None (Нет) в группе Icicle (Диаграмма накопления). Щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Hierarchical Cluster Analysis (Иерархический кластерный анализ).
6. Щелкните на кнопке Method (Метод), чтобы открыть диалоговое окно Hierarchical Cluster Analysis: Method (Иерархический кластерный анализ: Метод), показанное на рис. 22.4.
7. В списке Cluster Method (Метод кластеризации) оставьте выбранным пункт Between-groups linkage (Межгрупповое связывание), в списке Standardize (Стандартизация) выберите пункт Zscore ($z$-шкала) и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Hierarchical Cluster Analysis (Иерархический кластерный анализ).
8. Щелкните на кнопке Save (Сохранить), чтобы открыть диалоговое окно Hierarchical Cluster Analysis: Save New Variables (Иерархический кластерный анализ: Сохранение новых переменных), показанное на рис. 22.5.
9. Установите переключатель Single Solution (Заданное число кластеров), введите в расположенное рядом поле значение 3 и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Hierarchical Cluster Analysis (Иерархический кластерный анализ).
10. Щелкните на кнопке OK, чтобы открыть окно вывода.

В следующем примере проводится кластерный анализ, в котором вместо объектов участвуют переменные. Мы используем данные файла TestIQ.sav, содержащего 11 переменных и1, ..., и11. Обратите внимание на следующее обстоятельство. Обычно при группировании переменных исследователя интересует их взаимосвязь, а не их различие (сходство), как при группировании объектов. Исключением является случай, когда данные представляют собой оценки объектов экспертами — в этом случае строки соответствуют экспертам, а столбцы — оцениваемым объектам. Поскольку в нашем примере интерес представляют именно взаимосвязи между переменными и мы хотим сравнить результаты с факторным анализом, то в качестве меры близости целесообразно выбрать корреляцию. При этом корреляции надо учитывать по абсолютной величине, так как большие (по модулю) отрицательные

их величины так же свидетельствуют о связи, как и большие положительные. Все это необходимо иметь в виду, если речь идет о кластеризации переменных. Большинство остальных параметров команды оставим установленными по умолчанию; даже в стандартизации в данном случае нет необходимости, так как на величину корреляции не влияют единицы измерения переменных. Добавим лишь дендрограмму в выводимые результаты и исключим оттуда диаграмму накопления. Обратите внимание, поскольку для выполнения анализа необходим файл TestlQ.sav, а не cars.sav, который использовался в предыдущем примере, мы начинаем с шага 3а.

**Шаг 3а** Откройте файл данных, с которым вы намерены работать (в пашем случае это файл TestlQ.sav). Если он расположен в текущей папке, то выполните следующие действия:

1. Выберите в меню File (Файл) команду Open ▸ Data (Открытие ▸ Данные) или щелкните на кнопке Open File (Открытие файла) панели инструментов.
2. В открывшемся диалоговом окне дважды щелкните на имени TestlQ.sav или введите его с клавиатуры и щелкните па кнопке OK.

В меню Analyze (Анализ) выберите команду Classify ▸ Hierarchical Cluster (Классификация ▸ Иерархическая кластеризация). Откроется диалоговое окно Hierarchical Cluster Analysis (Иерархический кластерный анализ), показанное на рис. 22.1.

**Шаг 5а** После выполнения предыдущего шага должно быть открыто диалоговое окно Hierarchical Cluster Analysis (Иерархический кластерный анализ). Если вы уже успели поработать с этим окном, щелкните на кнопке Reset (Сброс).

1. В группе Cluster (Кластеризация) установите переключатель Variables (Переменные).
2. Нажмите кнопку мыши на переменной и1 и, не отпуская кнопки, перетащите указатель на переменную и11, затем кнопку мыши отпустите. В результате окажутся выделенными все 11 переменных и1, и2, ..., и11.
3. Щелкните на верхней кнопке со стрелкой, чтобы переместить выделенные переменные в список Variable(s) (Переменные).
4. Щелкните на кнопке Plots (Диаграммы), чтобы открыть диалоговое окно Hierarchical Cluster Analysis: Plots (Иерархический кластерный анализ: Диаграммы), показанное на рис. 22.3.
5. Установите флажок Dendrogram (Дендрограмма) и переключатель None (Нет) в группе Icicle (Диаграмма накопления). Щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Hierarchical Cluster Analysis (Иерархический кластерный анализ).
6. Щелкните на кнопке Method (Метод), чтобы открыть диалоговое окно Hierarchical Cluster Analysis: Method (Иерархический кластерный анализ: Метод), показанное на рис. 22.4.
7. В списке Interval (Интервал) выберите пункт Pearson correlation (Корреляция Пирсона), а в группе Transofm Measures (Преобразование значений) установите флажок Absolute values (Абсолютные значения). Щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Hierarchical Cluster Analysis (Иерархический кластерный анализ).
8. Щелкните на кнопке OK, чтобы открыть окно вывода.

Эта процедура реализует кластерный анализ 11 показателей теста интеллекта. В выводимые данные включается такая информации о переменных, как число корректных значений объектов, число утерянных значений и т. п. Затем выводятся последовательность слияния (agglomeration schedule) и горизонтальная дендрограмма. Рекомендуем сравнить полученные результаты с результатами факторного анализа в главе 20.

После выполнения шага 5 программа автоматически активизирует окно вывода. Для просмотра результатов вы при необходимости можете воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

# Печать результатов и выход из программы

Ниже описана типичная процедура печати результатов статистического анализа (или нескольких анализов). После выполнения шага 5 должно быть открыто окно вывода.

Шаг 6 В окне вывода укажите фрагменты, выводимые на печать (см. раздел «Окно вывода» в главе 2), в меню File (Файл) выберите команду Print (Печать), при необходимости задайте параметры печати и щелкните на кнопке OK.

Последнее, что необходимо сделать после завершения исследования и печати результатов, — это выйти из программы SPSS.

Шаг 7 Для выхода из программы в меню File (Файл) выберите команду Exit (Выход).

Иногда после выполнения команды Exit (Выход) на экране могут появляться небольшие диалоговые окна с вопросом о необходимости сохранения сделанных в файлах изменений и кнопками, описывающими возможные варианты ответа. Для завершения работы просто щелкайте на соответствующих кнопках.

# Представление результатов

В этом разделе представлены фрагменты выводимых данных, сгенерированные программой при выполнении шагов 5 и 5а.

Файл с исходными данными для нашего первого примера является небольшим, поэтому удобно включить исходные данные в вывод и использовать их при интерпретации результатов кластерного анализа. На рис. 22.6 приведена таблица с данными файла cars.sav. Строки соответствуют маркам автомобилей, столбцы — переменным.

Case Summaries

| | МАРКА | ЦЕНА | Т_СОСТ | ВОЗРАСТ | ПРОБЕГ |
|---|---|---|---|---|---|
| 1 | Audi | 16350 | 7 | 3 | 66000 |
| 2 | BMW | 14500 | 8 | 4 | 92500 |
| 3 | Buick | 8950 | 9 | 8 | 92500 |
| 4 | Chrysler | 8950 | 6 | 6 | 92500 |
| 5 | Dodge | 8450 | 5 | 6 | 92500 |
| 6 | Honda | 9850 | 7 | 4 | 118500 |
| 7 | Mazda | 12650 | 7 | 8 | 58000 |
| 8 | Mercedes | 17250 | 8 | 5 | 92500 |
| 9 | Mitsub. | 8950 | 3 | 5 | 136000 |
| 10 | Nissan | 9850 | 6 | 4 | 150000 |
| 11 | Pontiac | 8950 | 7 | 6 | 110000 |
| 12 | Saab | 14950 | 7 | 4 | 101000 |
| 13 | Toyota | 11700 | 8 | 7 | 57500 |
| 14 | VW | 8450 | 5 | 6 | 110000 |
| 15 | Volvo | 13100 | 8 | 4 | 75000 |

**Рис. 22.6.** Фрагмент окна вывода с данными файла cars.sav

Следующей частью выводимых результатов является показанная на рис. 22.7 таблица последовательности слияния. В этой таблице вторая колонка Cluster Combined (Объединенные кластеры) содержит первый (Cluster 1) и второй (Cluster 2) столбцы, которые соответствуют номерам кластеров, объединяемых на данном шаге. После объединения кластеру присваивается номер, соответствующий номеру в колонке Cluster 1. Так, на первом шаге объединяются объекты 5 и 14 и кластеру присваивается номер 5, далее этот кластер на шаге 3 объединяется с элементом 4 и новому кластеру присваивается номер 4 и т. д. Следующая колонка Coefficients (Коэффициент) содержит значение расстояния между кластерами, которые объединяются на данном шаге. Колонка Stage Cluster First Appears (Предыдущий шаг, на котором появлялся кластер) показывает, иа каком шаге до этого появлялись первый и второй из объединяемых кластеров. Последняя колонка Next Stage (Следующий шаг) показывает, иа каком шаге снова появится кластер, образованный на этом шаге.

Как видно по таблице, иа первом этапе происходит объединение в кластер пары объектов, расстояние между которыми является наименьшим. На втором этапе SPSS снова подсчитывает расстояния между объектами и объединяет в кластер пару наиболее близких объектов; при этом в результате может получиться либо один кластер из трех объектов, либо два кластера из двух объектов. Процесс слияния продолжается до тех пор, пока все объекты не попадут в один кластер. Попробуем объяснить результаты, полученные на этапах 1 и 13.

- *На этапе 1 происходит объединение объектов 5 и 14.* Расстояние между объектами равно 0,439. Ни один из двух объектов не принадлежит какому-либо кластеру, о чем свидетельствуют нули в столбцах Cluster 1 (Кластер 1) и Cluster 2 (Кластер 2) колонки Stage Cluster First Appears (Этап первого включения в кластер). Следующим этапом для данного кластера, судя по столбцу Next Stage (Следующий этап), является этап 3, на котором к кластеру присоединяется объект 4.
- *На этапе 13 происходит объединение кластеров, содержащих объекты 1 и 3.* Объект 1 был объединен с кластером, содержащим объект 2 на этапе 9, а объект

3 — с объектами 7 и 13 на этане 10. Расстояние между объединяемыми на этом этапе кластерами равно 9,656. Образованный на этом этапе кластер появляется далее на следующем шаге.

**Agglomeration Schedule**

| Stage | Cluster Combined | | Coefficients | Stage Cluster First Appears | | Next Stage |
|---|---|---|---|---|---|---|
| | Cluster 1 | Cluster 2 | | Cluster 1 | Cluster 2 | |
| 1 | 5 | 14 | .439 | 0 | 0 | 3 |
| 2 | 2 | 12 | .550 | 0 | 0 | 5 |
| 3 | 4 | 5 | .671 | 0 | 1 | 6 |
| 4 | 7 | 13 | .942 | 0 | 0 | 10 |
| 5 | 2 | 15 | 1.203 | 2 | 0 | 7 |
| в | 4 | 11 | 1.586 | 3 | 0 | 11 |
| 7 | 2 | в | 1.810 | 5 | 0 | 9 |
| в | 6 | 10 | 1.847 | 0 | 0 | 11 |
| 9 | 1 | 2 | 2.471 | 0 | 7 | 13 |
| 10 | 3 | 7 | 4.136 | 0 | 4 | 13 |
| 11 | 4 | в | 4.492 | 6 | 8 | 12 |
| 12 | 4 | 9 | 5.914 | 11 | 0 | 14 |
| 13 | 1 | 3 | 9.656 | 9 | 10 | 14 |
| 14 | 1 | 4 | 10.498 | 13 | 12 | 0 |

**Рис. 22.7.** Таблица последовательности слияния

По таблице последовательности слияния можно предварительно оценить число кластеров. Для этого необходимо проследить динамику увеличения различий по шагам кластеризации и определить шаг, па котором отмечается резкое возрастание различий. Оптимальному числу классов соответствует разность между числом объектов и порядковым номером шага, на котором был обнаружен заметный перепад различий. Так, в нашем примере резкое возрастание различий обнаруживается при переходе от шага 12 к шагу 13. Следовательно, наиболее оптимальное количество кластеров должно быть получено на шаге 12 или 13. Это количество равно численности объектов минус номер шага, то есть 15 - 12 = 3 или 15 - 13 = 2, то есть 3 или 2 кластера. Выбор того или иного решения зависит от конкретной ситуации.

Дендрограмма представляет процесс кластеризации в форме древовидной структуры (рис. 22.8).

Дендрограмма не только позволяет перейти к любому объекту на любом уровне кластеризации, но и дает возможность судить о том, каково расстояние между кластерами или объектами на каждом из уровней. Числа от 0 до 25 являются условной шкалой этих расстояний; 0 соответствует наименьшему расстоянию на первом этапе, а 25 — наибольшему расстоянию на последнем этапе.

На дендограмме любое решение характеризуется вертикальной линией, число точек пересечения которой с деревом соответствует количеству кластеров на текущем этапе. Так, для нашего примера эту линию следует расположить на уровне 15-20: между шагами 12 и 13 кластеризации. В этом случае получается 3 кластера. Для того чтобы установить состав каждого кластера, необходимо вернуться к корням

дерева и выяснить соответствующие номера объектов. Напомним, что в результате выполнения шага 5 в файле cars.sav появляется новая переменная, определяющая принадлежность каждого объекта к одному из трех кластеров.

Dendrogram using Average Linkage (Between Groups)

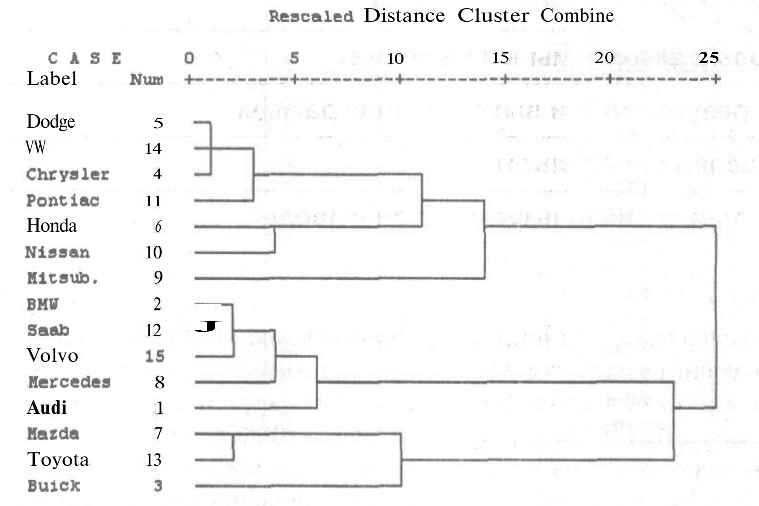



**Рис. 22.8.** Дендрограмма, полученная в результате выполнения шага 5

В заключение на рис. 22.9 приведем дендрограмму, полученную в результате выполнения шага 5а (кластеризация 11 переменных).

```
Dendrogram using Average Linkage (Between Groups)

                    Rescaled Distance Cluster Combine

  C A S E      0         5        10        15        20        25
 Label   Num   +---------+---------+---------+---------+---------+

 И4       4
 И8       8
 И7       7
 И9       9
 И2       2
 И10     10
 И5       5
 И6       6
 И1       1
 И11     11
 И3       3
```

**Рис. 22.9.** Дендрограмма, полученная в результате выполнения шага 5а

Дендрограмма показывает, что в результате кластеризации переменные группируются в три кластера, состав которых идентичен факторам, полученным в отношении тех же данных при факторном анализе (см. главу 20). Отметим, что кластерный анализ проще, а его результаты нагляднее.

# 23 Дискриминантный анализ



Дискриминантный анализ позволяет предсказать принадлежность объектов к двум или более непересекающимся группам. Исходными данными для дискриминантного анализа является множество объектов, разделенных на группы так, что каждый объект может быть отнесен только к одной группе. Допускается при этом, что некоторые объекты не относятся ни к какой группе (являются «неизвестными»). Для каждого из объектов имеются данные по ряду количественных переменных. Такие неременные называются дискриминантными переменными, или предикторами. Задачами дискриминантного анализа является определение:

- решающих правил, позволяющих по значениям дискриминантных переменных (предикторов) отнести каждый объект (в том числе и «неизвестный») к одной из известных групп;
- «веса» каждой дискриминантной переменной для разделения объектов на группы.

Существует множество ситуаций, в которых было бы весьма желательно вычислить вероятность того или иного исхода в зависимости от совокупности измеряемых переменных, например подходит ли соискатель работы на ту или иную должность, страдает ли психически больной человек шизофренией или психозом, вернется ли заключенный в тюрьму или к нормальной жизни после выхода на свободу, какие факторы влияют на увеличение риска пациента получить сердечный приступ и т. п. Во всех перечисленных ситуациях есть две общие черты: во-первых, для многих субъектов есть информация об их принадлежности к той или иной группе; во-вторых, о каждом субъекте имеется дополнительная информация для создания формулы, которая позволит спрогнозировать принадлежность субъекта к той или иной группе.

Дискриминантный анализ имеет определенное сходство с кластерным анализом; сходство заключается в том, что исследователь в обоих случаях ставит перед собой цель разделить совокупность объектов (а не переменных) на несколько

более мелких групп. Тем не менее процесс классификации в двух видах анализа принципиально различен. В кластерном анализе объекты классифицируются на основе их различия без какой-либо предварительной информации о количестве и составе классов. В дискриминантном анализе количество и состав классов изначально задан, и основная задача заключается в определении того, насколько точно можно предсказать принадлежность объектов к классам при помощи данного набора дискриминантных переменных (предикторов).

Дискриминантный анализ представляет собой альтернативу множественного регрессионного анализа (см. главу 18) для случая, когда зависимая переменная представляет собой не количественную (номинативную) переменную. При этом Дискриминантный анализ решает, по сути, те же задачи, что и множественный регрессионный анализ: предсказание значений «зависимой» перемененной (в данном случае категорий номинативного признака) и определение того, какие «независимые» переменные лучше всего подходят для такого предсказания. Дискриминантный анализ основан на составлении уравнения регрессии (см. главы 17 и 18), использующего номинативную зависимую переменную (обратите внимание на то, что она не является количественной, как в случае регрессионного анализа). Уравнение регрессии составляется на основе тех объектов, о которых известна групповая принадлежность, что позволяет максимально точно подобрать его коэффициенты. После того как уравнение регрессии получено, его можно использовать для группировки интересующих нас объектов в целях прогнозирования. Команда дискриминантного анализа весьма непроста и требует настройки множества параметров, описание большинства из которых лежит за рамками темы данной книги. Тем не менее при необходимости вы можете обратиться за дополнительной информацией к руководству пользователя SPSS.

Как и для большинства сложных статистических операций, параметры дискриминантного анализа в основном определяются особенностями данных, а также экспериментаторскими стремлениями исследователя. Как всегда, мы рассмотрим пример (па этот раз единственный) проведения дискриминантного анализа в разделе пошаговых процедур, а раздел «Представление результатов» посвятим интерпретации выводимых данных.

Для демонстрации дискриминантного анализа мы рассмотрим пример прогнозирования успешности обучения на основе предварительного тестирования. Файл class.sav содержит данные о 46 учащихся (объекты с 1 по 46), закончивших курс обучения, в отношении которых известны оценки успешности обучения — для этого используется переменная оценка (1 — низкая, 2 — высокая). Кроме того, в файл включены данные предварительного тестирования этих учащихся до начала обучения (13 переменных):

- и1...и11 — 11 показателей теста интеллекта;
- э_и — показатель экстраверсии по тесту Г. Айзенка (H. Eysenck);
- н — показатель нейротизма по тесту Г. Айзенка.

Еще для 10 претендентов на курс обучения (объекты с 47 по 56) известны лишь результаты их предварительного тестирования (13 перечисленных переменных).

Значения переменной оценка для них, разумеется, неизвестны, и в файле данных им соответствуют пустые ячейки. В процессе дискриминантного анализа мы, в частности, попытаемся спрогнозировать успешность обучения этих 10 претендентов в предположении, что выборки закончивших обучение и претендентов идентичны.

# Этапы дискриминантного анализа

Дискриминантный анализ состоит из трех основных этапов.

1. *Выбор переменных-предикторов.* Исследователь использует свои теоретические знания, практический опыт, догадки и т. п. для того, чтобы составить список переменных, которые могут повлиять на результат группировки (переменную-критерий). В рассматриваемом файле помимо переменной-критерия (оценка) содержится 13 переменных, характеризующих каждого учащегося; это позволяет нам сделать все 13 переменных предикторами и включить их в уравнение регрессии. Если бы число переменных было велико (например, несколько сотен), то было бы невозможно применить дискриминантный анализ ко всем переменным одновременно. Это обусловлено как концептуальными причинами (коллинеарность переменных, потеря степеней свободы и т. п.), так и практическими ограничениями (недостаточный объем оперативной памяти компьютера). Обычно на начальном этапе дискриминантного анализа для предикторов формируется корреляционная матрица. В данном контексте она имеет особый смысл, называется *общей внутригрупповой корреляционной матрицей* и содержит средние коэффициенты корреляции для двух или более корреляционных матриц (каждая для одной группы). Помимо общей внутригрупповой корреляционной матрицы можно также вычислить ковариационные матрицы для отдельных групп, для всей выборки либо общую внутригрупповую ковариационную матрицу. Нередко исследователи применяют серию $t$-критериев между двумя группами для каждой переменной либо однофакторный дисперсионный анализ, если число групп оказывается больше двух. Поскольку целью дискриминантного анализа является составление наилучшего уравнения регрессии, дополнительный анализ исходных данных никогда не является лишним. Так, в результате применения $t$-критериев для данных нашего примера были найдены значимые различия между двумя уровнями переменной оценка для 8 из 13 предикторов. Мы рассмотрим один из наиболее распространенных вариантов дискриминантного анализа, при проведении которого программа автоматически исключает несущественные для предсказания предикторы, но по критериям, которые устанавливает сам пользователь.

2. *Выбор параметров.* В этой главе будет продемонстрирован один из методов дискриминантного анализа. По умолчанию программа реализует метод, который основан на *принудительном включении* в регрессионное уравнение всех предикторов, указанных исследователем. В другом варианте используется метод Уилкса (Wilks), относящийся к категории *пошаговых методов* и основанный

на минимизации коэффициента Уилкса *(К)* после включения в уравнение регрессии каждого нового предиктора. Так же как и в случае множественного регрессионного анализа, существует критерий для включения предикторов в уравнение регрессии (по умолчанию таким критерием является $F > 3{,}84$) и критерий для исключения предикторов из уравнения регрессии (по умолчанию $F < 2{,}71$). Коэффициент Х представляет собой отношение внутригрупповой суммы квадратов к общей сумме квадратов и характеризует долю влияния предиктора на дисперсию критерия. Со значением Х связаны величины $F$ и $p$, характеризующие его значимость. Более полное описание вы можете найти в разделе «Представление результатов».

Какой же из двух методов предпочтительнее? Как показывает практика, зачастую компьютер справляется с составлением уравнения регрессии лучше, чем исследователь, задающий список предикторов вручную. Однако встречаются ситуации, когда полезней ограничить самостоятельность компьютера. Например, если провести дискриминантный анализ для наших данных с включением всех переменных, то неверно классифицированы будут 5 объектов из 46. Той же точности прогноза можно достичь всего с 7 предикторами, если выбрать пошаговый метод с установками, отличающимися от принятых по умолчанию. В то же время, если использовать пошаговый метод с установками по умолчанию, оставляющий только 3 предиктора, количество неверно сгруппированных объектов увеличится до 9. Помимо рассмотренных программа SPSS располагает и другими методами выбора предикторов, однако их описание выходит за рамки темы данной книги, и при необходимости мы рекомендуем вам обратиться к руководству пользователя SPSS.

3. *Интерпретация результатов.* Целью дискриминантного анализа является составление уравнения регрессии с использованием выборки, для которой известны значения и предикторов, и критерия. Это уравнение позволяет по известным значениям предикторов определить неизвестные значения критерия для другой выборки. Разумеется, точность рассчитываемых значений критерия для второй выборки в общем случае не выше, чем для исходной. Так, в нашем примере регрессионное уравнение обеспечило около 90 % корректных результатов для той выборки, с помощью которой оно было создано. Соответственно, точность предсказания успешности обучения для 10 претендентов может достигать 90 % лишь в том случае, если выборка претендентов совершенно идентична тем 46 учащимся, данные для которых послужили основой для прогноза.

# Пошаговые алгоритмы вычислений

Перед началом дискриминантного анализа выполняются три подготовительных шага.

**Шаг 1** Создайте новый файл данных или подготовьте существующий. Все необходимые действия описаны в главе 3.

Шаг 23 Запустите программу SPSS при помощи значка на рабочем столе или команды Пуск ▸ Программы ▸ SPSS for Windows ▸ SPSS 11.5 for Windows (Start ▸ Programs ▸ SPSS for Windows ▸ SPSS 11.5 for Windows) главного меню Windows. В открывшемся после запуска программы диалоговом окне SPSS for Windows щелкните на кнопке Cancel (Отмена).

После выполнения этого шага на экране появится окно редактора данных SPSS.

Шаг 3 Откройте файл данных, с которым вы намерены работать (в нашем случае это файл class.sav). Если он расположен в текущей папке, то выполните следующие действия:

1. Выберите в меню File (Файл) команду Open ▸ Data (Открытие ▸ Данные) или щелкните на кнопке Open File (Открытие файла) панели инструментов.
2. В открывшемся диалоговом окне дважды щелкните на имени class.sav или введите его с клавиатуры и щелкните на кнопке OK.

Независимо от того, открыта программа SPSS только что или какие-то процедуры уже выполнялись, в верхней части главного окна должна присутствовать строка меню (она показана ниже). Пока строка меню присутствует на экране, доступны все команды анализа данных. При этом окно с данными видеть не обязательно.

File Edit View Data Transform Analyze Graphs Utilities Window Help

При работе со сводными таблицами или при редактировании диаграмм некоторые пункты строки меню могут исчезать или меняться. Чтобы вернуться к главному окну и стандартной строке меню, щелкните на кнопке свертывания или восстановления текущего окна.

После завершения шага 3 на экране должно присутствовать окно редактора данных.

Шаг 4 В меню Analyze (Анализ) выберите команду Classify ▸ Discriminant (Классификация ▸ Дискриминантный анализ). На экране появится диалоговое окно Discriminant Analysis (Дискриминантный анализ), показанное на рис. 23.1.

Помимо таких привычных элементов интерфейса, как список доступных переменных слева и пять стандартных кнопок справа, диалоговое окно Discriminant Analysis (Дискриминантный анализ) содержит четыре дополнительных элемента.

- Поле Grouping Variable (Группирующая переменная) предназначено для задания единственной зависимой переменной.
- В список Independents (Независимые переменные) включается любое число переменных, участвующих в дискриминантном анализе в качестве предикторов.
- Переключатели Enter independents together (Вводить независимые переменные вместе) и Use stepwise method (Использовать пошаговый метод) позволяют выбрать один из двух вариантов дискриминантного анализа.

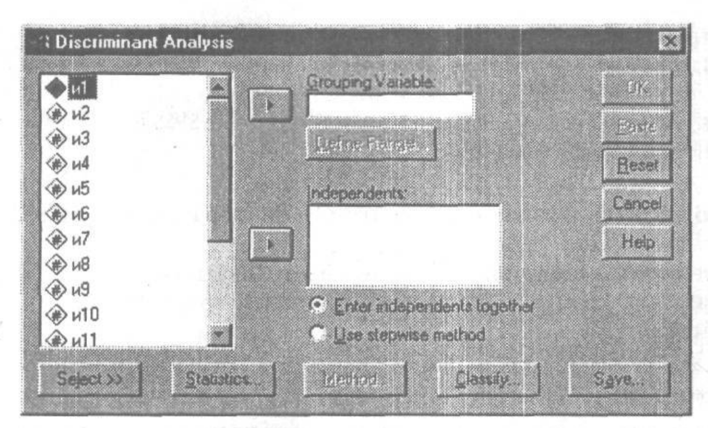



Рис. 23.1. Диалоговое окно Discriminant Analysis

- Пять кнопок управления дополнительными параметрами в нижней части диалогового окна предназначены для открытия дополнительных диалоговых окон. Далее будут показаны только три диалоговых окна, поскольку назначение элементов интерфейса, доступ к которым открывают кнопки Select (Выбор) и Save (Сохранение), достаточно очевидно.

Назначение кнопки Select (Выбор) примерно такое же, как у кнопки If (Если), упоминавшейся в главе 4. Появляющееся при щелчке на кнопке Select (Выбор) поле Selection Variable (Переменная для выбора) позволяет указать имя переменной и ее уровень. Если, к примеру, поместить в поле Selection Variable (Переменная для выбора) имя пол, щелкнуть на кнопке Value (Значение) и ввести значение 1, то в дискриминантном анализе будут участвовать только те объекты, для которых значения переменной пол равны 1.

. Кнопка Save (Сохранение) позволяет сохранять в качестве новых переменных следующие величины для каждого объекта (в том числе «неизвестного»):

- прогнозируемый номер группы;
- оценки дискриминантных функций;
- вероятность принадлежности к каждой группе.

Важными элементами интерфейса диалогового окна Discriminant Analysis (Дискриминантный анализ) являются элементы управления уровнями зависимой переменной. Хотя в нашем примере дискриминантный анализ используется для разделения объектов на две группы, количество групп может быть и больше. После ввода имени зависимой переменной в поле Grouping Variable (Группирующая переменная) вы можете щелкнуть на кнопке Define Range (Задать диапазон), в результате откроется диалоговое окно Discriminant Analysis: Define Range (Дискриминантный анализ: Задание диапазона), представленное на рис. 23.2. Поскольку переменная оценка, используемая в нашем примере как зависимая, имеет лишь два уровня (1 и 2),

следует указать их в полях Minumum (Минимум) и Maximum (Максимум), а затем щелкнуть на кнопке Continue (Продолжить). Если число уровней группирующей переменной больше двух, описанная операция позволяет задать любой диапазон уровней зависимой переменной.

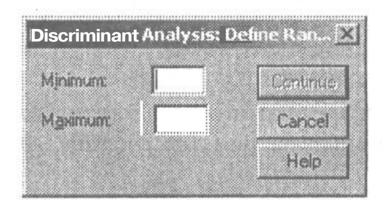



**Рис. 23.2.** Диалоговое окно Discriminant Analysis: Define Range

Щелчок на кнопке Statistics (Статистики) в окне Discriminant Analysis (Дискриминантный анализ) открывает диалоговое окно Discriminant Analysis: Statistics (Дискриминантный анализ: Статистики), представленное на рис. 23.3. Поскольку практически все элементы интерфейса этого окна активно используются исследователями, далее приведено описание каждого из них. Как правило, с помощью этих элементов исследователи выясняют наличие зависимостей между предикторами перед началом дискриминантного анализа.

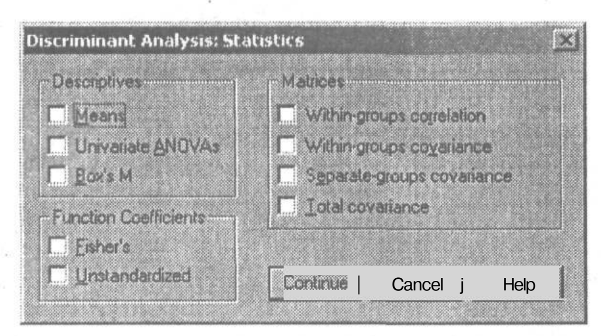



**Рис. 23.3.** Диалоговое окно Discriminant Analysis: Statistics

Флажки в группе Descriptives (Описательные статистики):

- Means (Средние) — средние значения и стандартные отклонения для каждой переменной каждой группы (в данном случае двух групп, образованных уровнями переменной оценка) и всей выборки в целом;
- Univariate ANOVAs (Однофакторный дисперсионный анализ) — если число уровней зависимой переменной равно 2, то это $t$-критерии для сравнения между собой двух средних значений для групп по каждой переменной;
- Box's M (M Бокса) — критерий равенства ковариационных матриц для уровней зависимой переменной. *M* Бокса является критерием многомерной нормальности исходных данных.

Флажки в группе Function Coefficients (Коэффициенты функций):

- Fisher's (Фишер) — канонические коэффициенты уравнения регрессии;
- Unstandardized (Нестандартизированные коэффициенты) — нестандартизованные коэффициенты уравнения регрессии, вычисленные для исходных значений предикторов.

Флажки в группе Matrices (Матрицы):

- Within-groups correlation (Корреляции внутри групп) — матрица, состоящая из соответствующих средних значений корреляционных матриц для уровней зависимой переменной;
- Within-groups covariance (Корреляции между группами) — матрица, аналогичная предыдущей с той разницей, что вместо корреляций используются ковариации;
- Separate-groups covariance (Ковариация отдельных групп) — отдельные ковариационные матрицы для каждого уровня зависимой переменной;
- Total covariance (Полная ковариация) — ковариационная матрица для всей выборки.

Щелчок на кнопке Method (Метод) в окне Discriminant Analysis (Дискриминантный анализ) приводит к появлению диалогового окна Discriminant Analysis: Stepwise Method (Дискриминантный анализ: Пошаговый метод), представленного на рис. 23.4. Обратите внимание, что кнопка Method (Метод) доступна лишь в том случае, если установлен переключатель Use stepwise method (Использовать пошаговый метод).

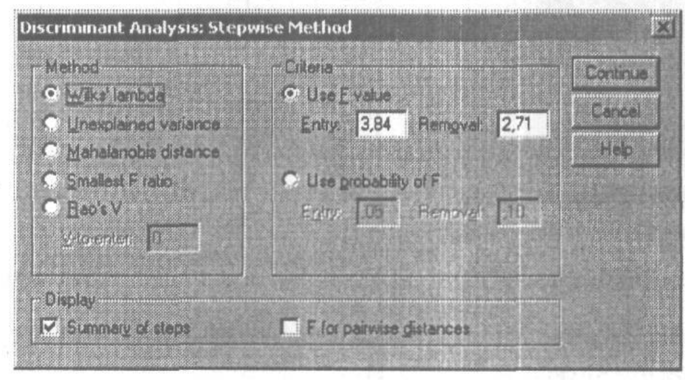



**Рис. 23.4.** Диалоговое окно Discriminant Analysis: Stepwise Method

В этом окне при помощи группы переключателей Method (Метод) можно выбрать один из пошаговых методов составления дискриминантного уравнения, в области Criteria (Критерии) задать критерии для включения в дискриминантное уравнение и исключения из него предикторов, а также включить в вывод желаемые величины при помощи флажков в группе Display (Отображать). Из переключателей выбора метода составления уравнения для нас представляет интерес лишь переключатель Wilks' lambda (Лямбда Уилкса). Соответствующий метод основан

на минимизации коэффициента Уилкса (λ) после включения в уравнение регрессии каждого нового предиктора. Более подробное описание приведено в разделе «Представление результатов».

По умолчанию критериями для включения и исключения предикторов являются пороговые значения *F*-критерия. Как правило, значению $F = 3{,}84$ соответствует величина уровня значимости *p*, приблизительно равная 0,05, а значению $F = 2{,}71$ – величина *p*, приблизительно равная 0,1. Исследователи нередко предпочитают использовать значения по умолчанию, нежели задавать собственные.

Флажок Summary of steps (Итоги по шагам) в группе Display (Отображать) управляет пошаговым выводом основных результатов дискриминантного анализа, поэтому по умолчанию он установлен. Флажок F for pairwise distances (F для парных расстояний) относится только к методу Mahalanobis distance (Расстояние Махаланобиса).

Щелчок на кнопке Classify (Классификация) в окне Discriminant Analysis (Дискриминантный анализ) открывает диалоговое окно Discriminant Analysis: Classification (Дискриминантный анализ: Классификация), представленное на рис. 23.5. Большая часть элементов интерфейса этого окна активно используется исследователями и описана далее.

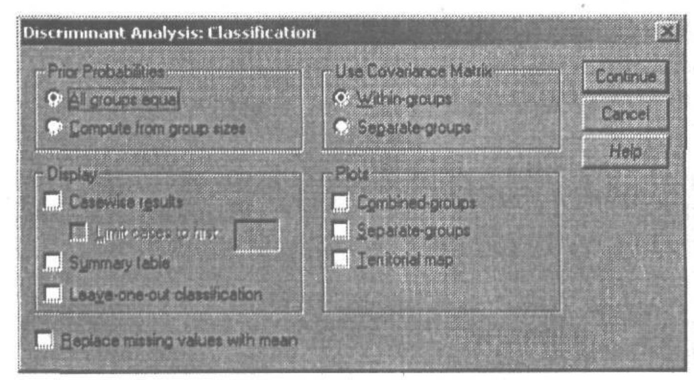



**Рис. 23.5.** Диалоговое окно Discriminant Analysis: Classification

В группе Prior Probabilities (Априорные вероятности) находятся два переключателя:

- All groups equal (Равные для всех групп) — вероятности принадлежности объекта к каждой из групп полагаются равными;
- Compute from group sizes (Пропорциональные размерам групп) — вероятности принадлежности объекта к каждой из групп пропорциональны размерам групп.

Флажки в группе Plots (Диаграммы):

- Combined-groups (Одна для всех групп) — в вывод включается гистограмма (если число групп равно 2) или диаграмма разброса (если число групп больше 2);
- Separate-groups (Отдельные для каждой группы) — в вывод включаются несколько диаграмм, каждая из которых соответствует одной группе;

- Territorial map (Карта территории) — этот флажок используется только в случаях, когда зависимая переменная имеет три или более уровней.

Два переключателя в группе Use Covariance Matrix (Использование ковариационной матрицы) позволяют выбрать ковариационную матрицу:

- Within-groups (Внутри групп) — этот переключатель установлен по умолчанию и классифицирует объекты при помощи общей внутригрупповой ковариационной матрицы;
- Separate-groups (Для отдельных групп) — объекты классифицируются при помощи отдельных ковариационных матриц для каждой группы.

Флажки в группе Display (Отображать) относятся к выводу результатов анализа:

- Casewise results (Результаты для объектов) — этот флажок полезно устанавливать, если файл данных не слишком велик: в результат включается список объектов, и для каждого объекта указывается фактическая группа, прогнозируемая группа, вероятность попадания в прогнозируемую группу и значения дискриминантных функций (подобный список приведен в разделе «Представление результатов»);
- Summary table (Итоговая таблица) — этот флажок позволяет включить в результат число и процент корректных и некорректных классификаций для каждой группы (в разделе «Представление результатов» имеются соответствующие результаты);
- Leave-one-out classification (Классификация по внешним данным) — при установке этого флажка каждый объект классифицируется с помощью функций, входными значениями для которых являются все объекты, в том числе те, для которых классификация неизвестна.

Название флажка Replace Missing Values with mean (Заменять пропущенные значения средним) вполне отражает его назначение.

В этой главе приведена единственная версия шага 5. Разумеется, вы можете провести дискриминантный анализ с параметрами по умолчанию, однако интересней провести анализ с применением некоторых из параметров, описанных в этой главе.

В следующем примере проводится дискриминантный анализ для зависимой переменной оценка, имеющей два уровня, и 13 предикторов. Предикторы добавляются в дискриминантное уравнение пошаговым методом (Уилкса) с установками, отличающимися от предлагаемых по умолчанию: для включения предикторов в уравнение служит значение $F = 1{,}25$, а для исключения предикторов — значение $F = 1{,}00$. Для анализа зависимости между предикторами выводятся все описательные статистики. Кроме того, мы включаем в вывод нестандартные коэффициенты дискриминантного уравнения, результаты для каждого объекта и итоговую таблицу.

Шаг 5 После выполнения предыдущего шага у вас должно быть открыто диалоговое окно Discriminant Analysis (Дискриминантный анализ), показанное на рис. 23.1. Если вы уже успели поработать с этим окном, щелкните на кнопке Reset (Сброс).

1. Щелкните сначала на переменной оценка, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в поле Grouping Variable (Группирующая переменная).
2. Щелкните на кнопке Define Range (Задать диапазон), чтобы открыть диалоговое окно Discriminant Analysis: Define Range (Дискриминантный анализ: Задание диапазона), показанное на рис. 23.2.
3. В поле Minimum (Минимум) введите значение 1, нажмите клавишу Tab, чтобы переместить фокус ввода в поле Maximum (Максимум), введите значение 2 и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Discriminant Analysis (Дискриминантный анализ).
4. Наведите указатель мыши на переменную и1, нажмите кнопку мыши и, не отпуская кнопки, перетащите указатель на последнюю переменную н, затем кнопку мыши отпустите. В результате окажутся выделенными все оставшиеся в списке переменные от и1 до н.
5. Щелкните на нижней кнопке со стрелкой, чтобы переместить выделенные переменные в список Independents (Независимые переменные), установите переключатель Use stepwise method (Использовать пошаговый метод) и щелкните на кнопке Statistics (Статистики), чтобы открыть диалоговое окно Discriminant Analysis: Statistics (Дискриминантный анализ: Статистики), показанное на рис. 23.3.
6. Установите флажки Means (Средние), Box's M (M Бокса), Univariate ANOVAs (Однофакторный дисперсионный анализ), Unstardardized (Нестандартизированные коэффициенты) и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Discriminant Analysis (Дискриминантный анализ).
7. Щелкните на кнопке Method (Метод), чтобы открыть диалоговое окно Discriminant Analysis: Stepwise Method (Дискриминантный анализ: Пошаговый метод), показанное на рис. 23.4.
8. Дважды нажмите клавишу Tab, чтобы переместить фокус ввода в поле Entry (Ввод), введите значение 1,125, нажмите клавишу Tab, чтобы переместить фокус ввода в поле Removal (Вывод), введите значение 1 и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Discriminant Analysis (Дискриминантный анализ).
9. Щелкните на кнопке Classify (Классификация), чтобы открыть диалоговое окно Discriminant Analysis: Classification (Дискриминантный анализ: Классификация), показанное на рис. 23.5.
10. Установите флажки Casewise results (Результаты для объектов), Summary table (Итоговая таблица) и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Discriminant Analysis (Дискриминантный анализ).
11. Щелкните на кнопке OK, чтобы открыть окно вывода.

После выполнения шага 5 программа автоматически активизирует окно вывода. Для просмотра результатов вы при необходимости можете воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

# Печать результатов и выход из программы

Ниже описана типичная процедура печати результатов статистического анализа (или нескольких анализов). После выполнения шага 5 должно быть открыто окно вывода.

Шаг 6 В окне вывода укажите фрагменты, выводимые на печать (см. раздел «Окно вывода» в главе 2), в меню File (Файл) выберите команду Print (Печать), при необходимости задайте параметры печати и щелкните на кнопке OK.

Последнее, что необходимо сделать после завершения исследования и печати результатов, — это выйти из программы SPSS.

Шаг 7 Для выхода из программы в меню File (Файл) выберите команду Exit (Выход).

Иногда после выполнения команды Exit (Выход) на экране могут появляться небольшие диалоговые окна с вопросом о необходимости сохранения сделанных в файлах изменений и кнопками, описывающими возможные варианты ответа. Для завершения работы просто щелкайте на соответствующих кнопках.

# Представление результатов

Далее приведены результаты, сгенерированные программой при выполнении шага 5. Как всегда, мы постарались представить данные, выведенные SPSS, в максимально удобной и понятной для вас форме. Из весьма объемных результатов вывода мы выбрали наиболее важные фрагменты, иллюстрирующие суть Дискриминантного анализа.

В таблице на рис. 23.6 для каждой переменной показаны средние значения на обоих уровнях (в обеих категориях) зависимой нереременной, а также общие средние значения.

В таблице на рис. 23.7 содержатся коэффициенты $\lambda$ Уилкса (строка Wilks' Lambda), $F$-критерии (строка F) и уровни значимости, характеризующие различия средних

значений для групп по каждой из переменных (строка Sig.). Наиболее важная для исследователя информация относится к значениям *F*-критерия и уровням значимости, поскольку именно по этим значениям можно судить, для каких переменных различие двух групп является значимым.

**Group Statistics**

| | оценка успешности | оо€едомлехност£ | скрSые фигурз | сропущехне сооє(в | 5ет в уме | сохятливость | из ключени изобрЕжений | ©хSогии | числовые ряды | умозакс5Ехиос | 2Siометрическое ссожехие | «оаучиОаниф слое | экSравЕрсис | хЕйротизS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Mean | низкая | 10.1 | 8.8 | 9.5 | 10.2 | 9.3 | 9.2 | 9.9 | 9.2 | 11.9 | 10.3 | 9.0 | 12.7 | 12.1 |
| | высокая | 11 5 | 10.9 | 10.8 | и q | 10.8 | 10.4 | 12 0 | 10.6 | 1? ? | 11 q | 10.3 | 13 S | 11.7 |
| | Total | 10.8 | 9R | 10.2 | 111 | 10.1 | 9.8 | 10.9 | qq | 12.0 | 111 | 9.7 | 13.1 | n q |

**Рис. 23.6.** Средние значения

**Tests of Equality of Group Means**

| | осведомленность | скрытые фигуры | пропущенные сло5 | счет в уме | понятливость | исключение изоБр[redacted]хsй | аналогии | числовые ряды | умозаключения | геометрическое сложен sE | заучивание слов | экстраверсия | хеSр[redacted] зм |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Wilks' Lambda | .90 | .85 | .93 | .81 | .84 | .93 | .87 | .92 | 1.00 | .91 | .91 | .99 | 1.00 |
| F | 4 86 | 7.75 | 3.27 | 10.56 | Я 13 | 3.06 | 6.81 | 4.08 | .16 | Л 37 | 4.48 | .63 | .14 |
| Sig. | .03 | .01 | .08 | .00 | .01 | .0Я | .01 | .05 | .69 | .04 | .04 | .43 | .71 |

**Рис. 23.7.** Коэффициенты лямбда Уилкса, F-критерии и уровни значимости

# Терминология, используемая при выводе

Ниже дана трактовка терминов, используемых программой в окне вывода и относящихся к средним значениям, λ Уилкса, *F*-критериям и уровням значимости.

- Mean (Средние значения) — средние значения для каждой категории (низкая, высокая) и для всех категорий (Total) переменной оценка.
- Wilks' Lambda (Лямбда Уилкса) — отношение внутригрупповой суммы квадратов к общей сумме квадратов (λ). Данный коэффициент характеризует долю

дисперсии оценок дискриминантной функции, которая не обусловлена различиями между группами, принимает значение 1 в случае, если средние значения для всех групп оказываются равными, и уменьшается с ростом разностей средних значений. Уровни значимости характеризуют вероятность того, что различия между группами являются случайными.

- F (*F*-критерий) — значения *F*-критерия такие же, как при однофакторном дисперсионном анализе, и равны квадрату *t*-критериев двух выборок.
- Sig. (Значимость) — уровни значимости *F*-критериев, равные вероятности того, что соответствующие различия являются случайными.

На рис. 23.8 представлены фрагменты выводимых результатов, связанные с регрессионным анализом и составлением дискриминантного уравнения. В трех таблицах содержится статистическая информация о переменных, вошедших и не вошедших в дискриминантное уравнение, а также порядок включения и исключения переменных в процессе составления дискриминантного уравнения. В ходе составления дискриминантного уравнения в него поочередно вводятся предикторы на основе заданного критерия включения (по умолчанию критерием является $F \geq 3{,}84$, в нашем случае — $F \geq 1{,}125$), а также исключаются из уравнения те предикторы, которые удовлетворяют критерию исключения (по умолчанию таким критерием является $F \leq 2{,}71$, в нашем случае — $F \leq 1{,}000$). Представленные таблицы являются завершающими фрагментами очень больших таблиц, которые генерирует программа в результате выполнения семи шагов заданного пошагового метода.

Обратите внимание, что все вошедшие в уравнение переменные после шага 7 имели достаточный уровень толерантности (выше 0,1) и значения *F*-критерия, превышающие пороговое значение 1,25. Переменные, не попавшие в дискриминантное уравнение, также имеют достаточную толерантность, однако значение *F*-критерия у них оказалось меньше 1,25. Последняя таблица иллюстрирует пошаговый процесс составления дискриминантного уравнения. Как можно видеть, шаги 1–7 приводят к последовательному введению в уравнение новых предикторов. Обращает на себя внимание тот факт, что в окончательный результат не вошли некоторые переменные, различия между группами для которых статистически достоверны. Напротив, в дискриминантное уравнение были включены переменные экстраверсия и умозаключения, различия между группами по которым статистически недостоверны. Это связано с тем, что при включении переменных в дискриминантное уравнение учитывается не только дискриминативная способность каждой переменной в отдельности, но и ее уникальный вклад в совокупности с остальными переменными.

Далее дана трактовка терминов, используемых программой в окне вывода и относящихся к составлению дискриминантного уравнения и регрессионному анализу.

- F to Enter (F для ввода) — минимальное значение *F*, при котором предиктор включается в дискриминантное уравнение.
- F to Remove (F для вывода) — максимальное значение *F*, при котором предиктор исключается из дискриминантного уравнения.

**Variables In the Analysis**

| Step | | Tolerance | F to Remove | Wilks' Lambda |
|---|---|---|---|---|
| 7 | счет в уме | .677 | 6.720 | .605 |
| | заучивание слов | .932 | 2.073 | .542 |
| | скрытые фигуры | .889 | 3.012 | .555 |
| | умозаключения | .549 | 5.832 | .593 |
| | аналогии | .600 | 3.006 | .555 |
| | понятливость | .855 | 2.525 | .549 |
| | экстраверсия | .886 | 1.797 | .539 |

**Variables Not In the Analysis**

| Step | | Tolerance | Min. Tolerance | F to Enter | Wilks' Lambda |
|---|---|---|---|---|---|
| 7 | осведомленность | .768 | .537 | .064 | .513 |
| | пропущенные слова | .662 | .509 | .014 | .514 |
| | исключение изображений | .790 | .542 | .016 | .514 |
| | числовые ряды | .555 | .522 | .096 | .513 |
| | геометрическое сложение | .519 | .519 | .000 | .514 |
| | нейротизм | .915 | .546 | 1.123 | .499 |

**Variables Entered/Removed**

| | | Wilks' Lambda | | |
|---|---|---|---|---|
| | | | Exact F | |
| Step | Entered | Statistic | Statistic | Sig. |
| 1 | счет в уме | .806 | 10.563 | .002 |
| 2 | заучивание слов | .708 | 8.849 | .001 |
| 3 | скрытые фигуры | .639 | 7.897 | .000 |
| 4 | умозаключения | .606 | 6.675 | .000 |
| 5 | аналогии | .570 | 6.043 | .000 |
| 6 | понятливость | .539 | 5.566 | .000 |
| 7 | экстраверсия | .514 | 5.125 | .000 |

**Рис. 23.8. Переменные**

- Tolerance (Толерантность) — мера линейной зависимости между одним предиктором и всеми остальными. Если величина толерантности окажется меньше 0,001, SPSS воспримет такой результат как наличие значительной линейной зависимости и не включит соответствующий предиктор в дискриминантное уравнение.
- Sig. (Значимость) — мера значимости влияния данного предиктора на дисперсию зависимой переменной.

Следующими представлены результаты теста многомерной нормальности данных (рис. 23.9).

**Box's Test of Equality of Covariance Matrices**

**Log Determinants**

| оценка успешности | Rank | Log Determinant |
|---|---|---|
| 1 | 7 | 9.653 |
| 2 | 7 | 11.017 |
| Pooled within-groups | 7 | 11.234 |

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

**Test Results**

| Box's M | F | | | |
|---|---|---|---|---|
| | Approx. | df1 | df2 | Sig. |
| 39.543 | 1.170 | 28 | 6746.134 | .246 |

Tests null hypothesis of equal population covariance matrices.

Рис. 23.9. Результаты теста многомерной нормальности данных

Ниже дана трактовка терминов, используемых программой в окне вывода и относящихся к результатам теста многомерной нормальности данных.

- Rank (Ранг) — размер ковариационной матрицы. Число 7 соответствует числу переменных в дискриминантном уравнении и говорит о том, что матрица имеет размер $7 \times 7$.
- Log determinant (Логарифмический определитель) — в этом столбце перечислены натуральные логарифмы определителей каждой из ковариационных матриц (двух в данном случае).
- Pooled within-groups (Общая внутригрупповая ковариационная матрица) — матрица, состоящая из средних значений ковариационных матриц, построенных для каждого уровня зависимой переменной.
- Box's M (M Бокса) — критерий многомерной нормальности, основанный на близости значений определителей ковариационных матриц.
- Approx (Приблизительная величина $F$-критерия) — величина, характеризующая близость значений определителей ковариационных матриц. Ее можно сравнить с $F$-критерием в дисперсионном анализе, когда межгрупповой разброс сравнивался с внутригрупповым.
- Sig. (Значимость) — если значимость превышает 0,05, данные характеризуются многомерной нормальностью. Если значимость равна или меньше 0,05, многомерная нормальность считается нарушенной, однако это не означает, что дискриминантный анализ неприменим к данным; более того, включение в уравнение переменной, чье распределение отличается от нормального, иногда может улучшить результат анализа. Отсутствие многомерной нормальности указывает лишь на то, что не все используемые переменные имеют нормальное распределение.

На рис. 23.10 и 23.11 приведены очередные таблицы, содержащие коэффициенты канонической дискриминантной функции, корреляции между каждым из предикторов и дискриминантной функцией, нестандартизированные коэффициенты, а также центроиды групп.

**Structure Matrix**

| Function | 3ет в ), а | ятлизоб6 , | окрытые фигур 3 2 | налогии та та | зачивание 5ов & | £ ометрическое оложение | осведомленность а | числовые ряды га | пропущенные 5оЕа а | исключение изо ажний а | хфйроти я | экстраверси к | у аключехжя 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | .504 | .442 | .432 | .405 | .328 | .326 | .290 | .259 | .258 | .245 | .181 | .123 | .063 |

**a. This variable not used in the analysis.**

**Рис. 23.10.** Структурная матрица

Далее дана трактовка терминов, используемых программой в окне вывода и относящихся к канонической дискриминантной функции и центроидам групп.

- Canonical Discriminant function Coefficients (Коэффициенты канонической дискриминантной функции) — список нестандартизованных (Unstandardized) коэффициентов и константа (Constant) дискриминантного уравнения. Это уравнение подобно линейному уравнению множественной регрессии. Значение функции для каждого объекта подсчитывается по этому уравнению.
- Function (Функция), Test of function (Тест функции) — значение 1 в этой ячейке говорит о том, что в процессе дискриминантного анализа была получена одна дискриминантная функция. Если бы зависимая переменная имела не 2, а 3 уровня, то было бы составлено две дискриминантные функции.
- Wilks' Lambda (Лямбда Уилкса) — отношение внутригрупповой суммы квадратов к общей сумме квадратов ($\lambda$).
- Chi-square (Хи-квадрат) — мера статистического отличия друг от друга двух уровней дискриминанта ($x^2$). Чем больше данное значение, тем сильнее отличие и тем лучше дискриминантная функция соответствует своему назначению.
- df (Число степеней свободы) — количество переменных, входящих в состав дискриминантного уравнения (7).
- Sig. (Значимость) — значимость, относящаяся к хи-квадрат.
- Eigenvalue (Собственное значение) — отношение межгрупповой сумы квадратов к внутригрупповой сумме квадратов. Чем больше данное значение, тем предпочтительнее для дискриминантного анализа составленная функция.

**Canonical Discriminant Function Coefficients**

| Function | скрып фигуры | счет в уме | понятливость | [illegible] | заключения | заучивание слов | экстраверсия | (Constant) |
|---|---|---|---|---|---|---|---|---|
| 1 | .162 | .382 | .214 | .185 | -.241 | .157 | .097 | -9.865 |

Unstandardized coefficients

**Functions at Group Centroids**

| оценка успешности | Function 1 |
|---|---|
| 1 | -.950 |
| 2 | .950 |

Unstandardized canonical discriminant functions evaluated at group means

**Eigenvalues**

| Function | Eigenvalue | % of Variance | Cumulative % | Canonical Correlation | Test of Function(s) | Wilks' Lambda | Chi-square | df | Sig |
|---|---|---|---|---|---|---|---|---|---|
| 1 | .944 | 100 | 100 | .697 | 1 | .514 | 26.93 | 7 | .000 |

**Рис. 23.11.** Результаты, относящиеся к канонической дискриминантной функции и центроидам групп

- % of variance (Процент дисперсии), Cumulative % (Накопленный процент дисперсии) — дискриминантная функция всегда вычисляется для равной 100 % дисперсии зависимой переменной.
- Structure Matrix (Структурная матрица) — структурная матрица содержит корреляции между значениями дискриминантной функции и каждой из переменных. Переменные упорядочены но абсолютной величине корреляций.
- Group centroids (Центроиды групп) — средние значения дискриминантной функции для двух групп. Более точно, центроид представляет собой значение функции, получаемое при подстановке в дискриминантное уравнение средних значений предикторов. Обратите внимание, что центроиды равны по абсолютной величине, но имеют разные знаки. Граничным значением для двух групп является ноль.

Таблица, представленная на рис. 23.12, содержит информацию о фактической и прогнозируемой группах для каждого объекта, а также значения дискриминантов. В целях экономии места мы привели значения лишь для 13 из 56 объектов. Для объектов, отмеченных двумя звездочками (**), фактическая и прогнозируемая группы не совпали. Всего таких объектов 5 из 46. Таким образом, точность классификации при данном наборе дискриминантных переменных составляет 89,13 % (41 из 46 правильных предсказаний в отношении «известных» объектов). В отношении последних 10 объектов, для которых принадлежность к группе не была известна, в таблице представлены результаты предсказания, полученные при помощи уравнения дискриминантной функции.

**Casewise Statistics**

| Case № | Actual Group | Highest Group | | | Second Highest Group | | Discriminant Scores |
|---|---|---|---|---|---|---|---|
| | | Predicted Group | P(D>d \| G=g) | P(G=g \| D=d) | Group | P(G=g \| D=d) | |
| 1 | 1 | 1 | .994 | .861 | 2 | .139 | -.958 |
| 2 | 2 | 2 | .853 | .896 | 1 | .104 | 1.135 |
| 3 | 1 | 1 | .516 | .639 | 2 | .361 | -.300 |
| 4 | 2 | 2 | .434 | .964 | 1 | .036 | 1.732 |
| 5 | 2 | 2 | .974 | .866 | 1 | .134 | .983 |
| 6 | 2 | 1** | .507 | .633 | 2 | .367 | -.287 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 32 | 2 | 1** | .100 | .993 | 2 | .007 | -2.594 |
| 33 | 2 | 2 | .375 | .971 | 1 | .029 | 1.838 |
| 34 | 1 | 1 | .497 | .626 | 2 | .374 | -.271 |
| 35 | 2 | 1** | .459 | .598 | 2 | .402 | -.210 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 46 | 1 | 1 | .644 | .717 | 2 | .283 | -.488 |
| 47 | ungrouped | 2 | .997 | .858 | 1 | .142 | .946 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 56 | ungrouped | 1 | .499 | .957 | 2 | .043 | -1.627 |

**. Misclassified case

**Рис. 23.12.** Данные о фактической и прогнозируемой группах для каждого объекта, а также значения дискриминантов

Далее дана трактовка терминов, используемых программой в окне вывода.

- Ungrouped (Несгруппированный объект) — объект, для которого заранее неизвестна принадлежность к группе.
- Actual group (Фактическая группа) — группа, которой принадлежит данный объект.
- Predicted Group (Прогнозируемая группа) — группа, вычисленная для объекта с помощью уравнения дискриминантной функции.

- P (D>d | G=g) — вероятность принадлежности объекта к группе (G) при данной величине дискриминантной функции (D).
- P (G=g | D=d) — вероятность наблюдаемого значения дискриминантной функции (D), если задана принадлежность объекта к группе (G).
- Highest Group (Вероятная группа) — группа, имеющая наибольшую прогнозируемую вероятность включения данного объекта.
- Second Highest Group (Вторая вероятная группа) — группа, имеющая вторую по величине вероятность (после прогнозируемой) включения данного объекта. Поскольку число групп в данном случае равно 2, то такая группа для каждого объекта определена «заранее».
- Discriminant Scores (Значения дискриминантной функции) — величины, получаемые при подстановке значений неременных объекта в уравнение дискриминантной функции.

# 24 Логистическая регрессия



Логистическая регрессия представляет собой расширение множественной регрессии, рассмотренной в главе 18, и отличается от последней тем, что в качестве зависимой переменной используется не количественная, а дихотомическая переменная, имеющая лишь два возможных значения[1]. Как правило, эти два значения символизируют принадлежность или непринадлежность объекта какой-либо группе, ответ типа «да» или «пет» и т. и.

Поскольку логистическая и множественная регрессии основаны на сходных принципах, мы будем исходить из того, что вы уже с ними знакомы (см. главу 18). Кратко смысл регрессионного анализа можно свести к нахождению аналитического выражения, наиболее адекватно отражающего связь между зависимой переменной и множеством независимых переменных. Главным отличием логистической регрессии от множественной является толкование уравнения регрессии. Если множественная регрессия позволяет прогнозировать количественное значение зависимой переменной (критерия) па основе известных значений независимых переменных (предикторов), то логистическая регрессия прогнозирует вероятность некоторого события, находящуюся в пределах от 0 до 1. Кроме того, при помощи индикаторной схемы кодирования допускается использование в качестве предикторов категориальных (номинативных) переменных. Более детально особенности логистической регрессии рассмотрены в разделе «Представление результатов», сейчас же для вас достаточно знать о том, что категориальный предиктор может быть представлен серией бинарных переменных — по одной на каждую категорию предиктора. Этим бинарным переменным присваиваются значения 1 или 0 в зависимости от того, к какой категории относится объект.

Математически суть логистической регрессии излагается па примере переменных из файла helpLR.sav. Содержимое этого файла примерно то же, что и файла help.sav, который использовался в главе 18. При помощи файла help.sav прогнозировалось

---

[1] Реализация логистической регрессии требует полной установки программы SPSS, включая модули Advanced Models (Дополнительные модели) и Regression Models (Регрессионные модели).

количественное значение переменной помощь в зависимости от ряда количественных предикторов (симпатия, проблема, эмпатия, польза и агрессия). В нашем случае мы будем прогнозировать не величину помощи, а мнение партнера о том, полезна или нет оказанная ему помощь. Как нетрудно понять, речь идет о дихотомической зависимой переменной, поскольку предполагается лишь два варианта ответа: «да» или «нет». Помимо того что зависимая переменная польза в файле helpLR.sav представлена в дихотомическом виде, в него добавлена еще одна категориальная переменная условия. Эта переменная обозначает одно из трех условий, в которых моделировалось оказание помощи: 1 — на работе, 2 — на улице в людном месте, 3 — в лесу, при встрече «один на один». Данные для файла helpLR.sav подбирались искусственно специально для того, чтобы проиллюстрировать рассматриваемый тип статистического анализа.

# Математическое описание логистической регрессии

С логистической регрессией связаны такие математические понятия, как вероятность, шанс и натуральный логарифм шанса. Вероятность — это ожидаемая относительная частота некоторого события. Шанс представляет собой отношение вероятности того, что событие произойдет, к вероятности того, что событие не произойдет. Так, если вероятность дождя равна 0,2, то вероятность отсутствия дождя равна 0,8, следовательно, шанс, что дождь все-таки прольется, равен 0,2/0,8 = 0,25. Обратите внимание, что шанс, в отличие от вероятности, не ограничен максимальным единичным значением; если, к примеру, вероятность дождя составляет не 0,2, а 0,8, то получаем шанс 0,8/0,2 = 4. Единичное значение шанса соответствует ситуации, когда вероятности появления и непоявления события равны.

Ключевым параметром логистической регрессии является *логит*. Логит равен натуральному логарифму шанса. Например, логит вероятности в 20 % равен −1,386...

Уравнение регрессии, используемое в главе 18, имеет следующий вид:

$$\text{помощь - } B_0 + B_1 \text{ x симпатия } + B_2 \text{ x агрессия } + B_3 \text{ x польза.}$$

Согласно этому уравнению величина оказываемой помощи равна сумме констанды ($B_0$) и значений трех переменных (отражающих симпатию, агрессивность и пользу), умноженных на соответствующие коэффициенты регрессии. Несмотря на то, что в логистическом анализе оценивается полезность или бесполезность помощи, а не ее величина, уравнение логистической регрессии похоже на уравнение множественной регрессии:

$$\ln\left[\frac{P_{\text{помощь}}}{1-P_{\text{помощь}}}\right] = B_0 + B_1x_1 + B_2x_2 + B_3x_3.$$

Здесь $P_{помощь}$ — вероятность оказания помощи, $x_1$, $x_2$ и $x_3$ — значения переменных симпатия, агрессия и польза.

Можно избавиться от натурального логарифма в левой части и преобразовать уравнение к виду:

$$\frac{P_{помощь}}{1-P_{помощь}} = e^{B_0} \times e^{B_1x_1} \times e^{B_2x_2} \times e^{B_3x_3}.$$

Это уравнение, в свою очередь, можно преобразовать к следующему:

$$P_{помощь} = \frac{1}{1+e^{-B_0} \times e^{-B_1x_1} \times e^{-B_2x_2} \times e^{-B_3x_3}},$$

Вероятно, подобные формулы не дают интуитивного представления о зависимости, которую они отражают, но на первых этапах это вполне нормально, поскольку интуиция всегда приходит с опытом. Тем не менее вам следует проявить максимум внимания к интерпретации модели логистической регрессии, поскольку выбор модели требует ее ясного теоретического понимания.

Поскольку данный раздел статистики весьма непрост для изучения, мы рекомендуем перед проведением логистического регрессионного анализа почитать дополнительную литературу.

# Пошаговые алгоритмы вычислений

Перед началом логистического регрессионного анализа необходимо выполнить три подготовительных шага.

**Шаг 1** Создайте новый файл данных или подготовьте существующий. Все необходимые действия описаны в главе 3.

**Шаг 2** Запустите программу SPSS при помощи значка на рабочем столе или команды Пуск ▸ Программы ▸ SPSS for Windows ▸ SPSS 11.5 for Windows (Start ▸ Programs ▸ SPSS for Windows ▸ SPSS 11.5 for Windows) главного меню Windows. В открывшемся после запуска программы диалоговом окне SPSS for Windows щелкните на кнопке Cancel (Отмена).

После выполнения этого шага на экране появится окно редактора данных SPSS.

**Шаг 3** Откройте файл данных, с которым вы намерены работать (в нашем случае это файл helpLR.sav). Если он расположен в текущей папке, то выполните следующие действия:

1. Выберите в меню File (Файл) команду Open > Data (Открытие ▸ Данные) или щелкните на кнопке Open File (Открытие файла) панели инструментов.
2. В открывшемся диалоговом окне дважды щелкните на имени helpLR.sav или введите его с клавиатуры и щелкните на кнопке OK.

Независимо от того, открыта программа SPSS только что или какие-то процедуры уже выполнялись, в верхней части главного окна должна присутствовать строка меню (она показана ниже). Пока строка меню присутствует на экране, доступны все команды анализа данных. При этом окно с данными видеть не обязательно.

File Edit View Data Transform Analyze Graphs Utilities Window Help

При работе со сводными таблицами или при редактировании диаграмм некоторые пункты строки меню могут исчезать или меняться. Чтобы вернуться к главному окну и стандартной строке меню, щелкните на кнопке свертывания или восстановления текущего окна.

После завершения шага 3 на экране должно присутствовать окно редактора данных.

**Шаг 4** В меню Analyze (Анализ) выберите команду Regression ▶ Binary Logistic (Регрессия ▶ Бинарная логистическая регрессия). На экране появится диалоговое окно Logistic Regression (Логистическая регрессия), показанное на рис. 24.1.

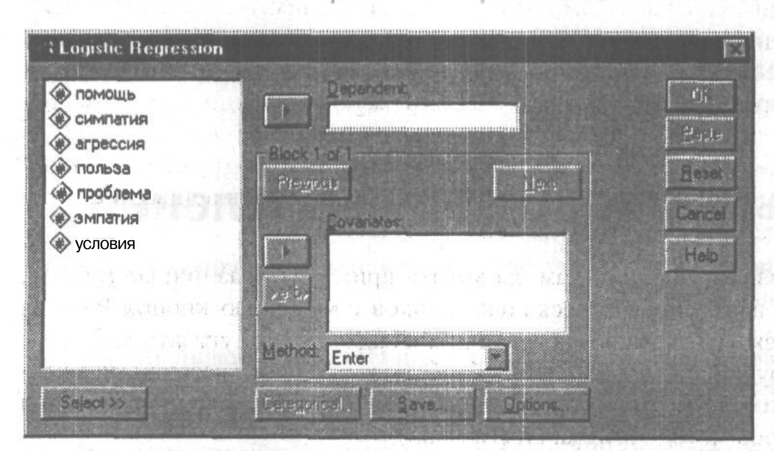



**Рис. 24.1.** Диалоговое окно Logistic Regression

В диалоговом окне Logistic Regression (Логистическая регрессия) вы можете ввести имена зависимой переменной и предикторов, а также задать другие параметры анализа. Как обычно, в левой части диалогового окна располагается список всех переменных файла данных. Справа от списка находится поле Dependent (Зависимая переменная); в это поле нужно поместить имя бинарной переменной, значения которой должны прогнозироваться в процессе анализа. В нашем примере в качестве зависимой будет использоваться переменная помощь, имеющая два уровня: 0 (нет) означает неоказанную помощь, а 1 (есть) — оказанную. Особенностью логистического анализа является то, что в процессе его проведения любые значения двух уровней зависимой переменной заменяются нулем и единицей. Это всегда происходит автоматически, и вам не нужно предпринимать никаких дополнительных действий для перекодирования переменной.



В центре окна Logistic Regression (Логистическая регрессия) находится список Covariates (Ковариаты), предназначенный для задания предикторов, что, как обычно, выполняется щелчком на кнопке со стрелкой.

Помимо задания обычных переменных вы можете указать способ их взаимодействия; для этого следует выделить те переменные, которые участвуют во взаимодействии, а затем щелкнуть на кнопке >a*Ь>. Однако при этом следует помнить, что подобные действия ведут к не вполне очевидным последствиям, и для успешной интерпретации результатов вы должны иметь о подобных действиях ясное концептуальное представление.

Раскрывающийся список Method (Метод) позволяет выбрать способ построения уравнения регрессии. Пункт Enter (Включение), выбранный по умолчанию, предполагает включение в регрессионное уравнение всех указанных предикторов независимо от того, имеют они значимое влияние на зависимую переменную или нет. Пункт Forward: LR (Прямой) предполагает пошаговое включение в уравнение предикторов, оказывающих наибольшее воздействие на зависимую переменную, до последнего предиктора, чье воздействие окажется значимым. Пункт Backward: LR (Обратный) предполагает обратный принцип: сначала в уравнение регрессии включаются все выбранные предикторы, а затем происходит пошаговое исключение тех предикторов, влияние которых на зависимую переменную оказывается недостаточным согласно установленному критерию; исключение проводится до тех пор, пока в уравнении останутся только те предикторы, которые удовлетворяют заданному критерию. Описанные три метода из представленных в списке Method (Метод) используются чаще других.

В случае если к разным предикторам вы хотите применить различные методы, следует разделить последние на несколько блоков с помощью кнопок Previous (Предыдущий) и Next (Следующий), а затем для каждого блока указать собственный метод. SPSS будет последовательно обрабатывать блоки указанными методами. В нашем примере мы не станем делить предикторы на блоки, а выберем для всех предикторов метод Forward: LR (Прямой).

Если в вашем анализе используются категориальные предикторы, то после задания всех предикторов в списке Covariates (Ковариаты) следует воспользоваться кнопкой Categorical (Категории). На экране появится диалоговое окно Logistic Regression: Define Categorical Variables (Логистическая регрессия: Определение.категориальных переменных), представленное на рис. 24.2.

В списке Covariates (Ковариаты) вы можете видеть имена всех предикторов, указанных вами в одноименном списке диалогового окна Logistic Regression (Логистическая регрессия). В нашем примере неременная условие является не количественной, а категориальной, поэтому ее следует выделить и переместить из списка Covariates (Ковариаты) в список Categorical Covariates (Категориальные ковариаты). Допускается применение нескольких способов обработки категориальных ковариат, каждый из которых задается при помощи списка Contrast (Контраст), находящегося в правом нижнем углу диалогового окна.

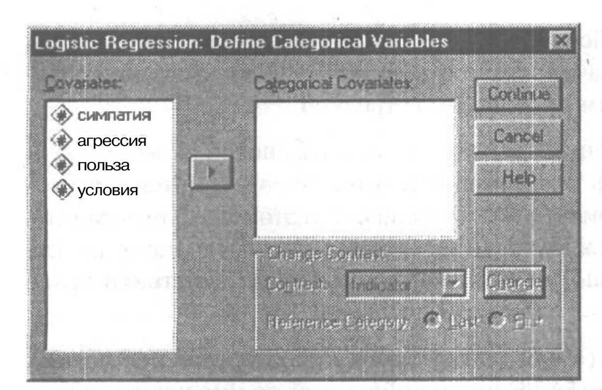



Рис. 24.2. Диалоговое окно Logistic Regression: Define Categorical Variables

Кнопка Options (Параметры) в диалоговом окне Logistic Regression (Логистическая регрессия) позволяет задать несколько дополнительных параметров, иногда оказывающихся весьма полезными на практике. При щелчке на ней открывается диалоговое окно Logistic Regression: Options (Логистическая регрессия: Параметры), представленное на рис. 24.3.

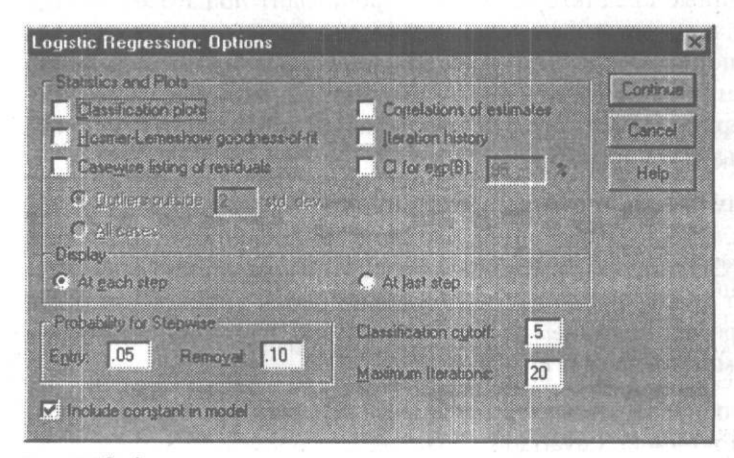



Рис. 24.3. Диалоговое окно Logistic Regression: Options

Флажок Classification Plots (Классификационная диаграмма) позволяет включить в выводимые результаты диаграмму, на которой для каждого объекта вместе с вероятностью прогноза указана его классификация, полученная при помощи регрессионного уравнения. Таким образом, диаграмма наглядно иллюстрирует, насколько адекватным оказалось составленное уравнение регрессии. Флажок Correlations of estimates (Корреляции оценок) указывает на необходимость построения корреляционной матрицы для всех переменных, включенных в регрессионное уравнение. В области Probability for Stepwise (Пошаговые вероятности) имеются поля Entry (Включение) и Removal (Исключение). В нервом указывается значение вероятности, служащее критерием включения предиктора в уравнение регрессии, а во втором -

значение вероятности, которое должен иметь каждый предиктор, исключаемый из состава уравнения регрессии (критерий исключения из уравнения регрессии). По умолчанию в иоле Entry (Включение) указано значение 0,05, а в поле Removal (Исключение) — значение 0,10. При желании исследователь может изменять их.

Кнопка Save (Сохранить) позволяет сохранить вычисленные значения вероятностей (Probabilities) и предсказанные значения зависимой переменной (Group membership) как новые переменные для объектов в файле данных. Это зачастую полезно для сравнения фактической и предсказанной группировок для отдельных объектов. В практическом плане интерес могут представлять предсказания для тех объектов, для которых фактические значения зависимой нереременной неизвестны.

В следующем примере проводится логистический регрессионный анализ, в котором в качестве зависимой переменной выступает переменная помощь, а в качестве предикторов — переменные симпатия, агрессия, польза и условие.

**Шаг 5** После выполнения предыдущего шага у вас должно быть открыто диалоговое окно Logistic Regression (Логистическая регрессия), показанное на рис. 24.1. Если вы уже успели поработать с этим окном, щелкните на кнопке Reset (Сброс).

1. Щелкните сначала на переменной помощь, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в список Dependent (Зависимая переменная).
2. Щелкните сначала на переменной симпатия, чтобы выделить ее, а затем — на нижней кнопке со стрелкой, чтобы переместить переменную в список Covariates (Ковариаты).
3. Повторите предыдущее действие для переменных агрессия, польза и условие.
4. В списке Method (Метод) выберите пункт Forward: LR (Прямой) и щелкните на кнопке Categorical (Категории), чтобы открыть диалоговое окно Logistic Regression: Define Categorical Variables (Логистическая регрессия: Определение категориальных переменных), показанное на рис. 24.2.
5. Щелкните сначала на переменной условие, чтобы выделить ее, а затем — на кнопке со стрелкой, чтобы переместить переменную в список Categorical Covariates (Категориальные ковариаты).
6. В списке Contrasts (Контрасты) выберите пункт Deviation (Отклонение) и для подтверждения выбора щелкните на кнопке Change (Изменить). Щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Logistic Regression (Логистическая регрессия).
7. Щелкните на кнопке Options (Параметры), чтобы открыть диалоговое окно Logistic Regression: Options (Логистическая регрессия: Параметры), показанное на рис. 24.3.
8. Установите флажки Classification Plots (Классификационная диаграмма) и Correlations of estimates (Корреляции оценок), а затем щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Logistic Regression (Логистическая регрессия).
9. Щелкните на кнопке OK, чтобы открыть окно вывода.

После выполнения шага 5 программа автоматически активизирует окно вывода. Для просмотра результатов вы при необходимости можете воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

# Печать результатов и выход из программы

Ниже описана типичная процедура печати результатов статистического анализа (или нескольких анализов). После выполнения шага 5 должно быть открыто окно вывода.



Шаг 6 В окне вывода укажите фрагменты, выводимые на печать (см. раздел «Окно вывода» в главе 2), в меню File (Файл) выберите команду Print (Печать), при необходимости задайте параметры печати и щелкните на кнопке OK.

Последнее, что необходимо сделать после завершения исследования и печати результатов, — это выйти из программы SPSS.

Шаг 7 Для выхода из программы в меню File (Файл) выберите команду Exit (Выход).

Иногда после выполнения команды Exit (Выход) на экране могут появляться небольшие диалоговые окна с вопросом о необходимости сохранения сделанных в файлах изменений и кнопками, описывающими возможные варианты ответа. Для завершения работы просто щелкайте на соответствующих кнопках.

# Представление результатов

Далее приведены результаты, сгенерированные программой при выполнении шага 5.

Поскольку методы Forward: LR (Прямой) и Backward: LR (Обратный) являются пошаговыми, при их использовании выводится несколько таблиц, характеризующих каждый шаг и позволяющих наглядно представить себе процесс формирования уравнения регрессии. Здесь мы ограничились лишь итоговой таблицей.

## Кодирование категориальных переменных

Если в вашем анализе использовались категориальные ковариаты и контрасты, то в начало выводимых результатов будет включена таблица (рис. 24.4), показывающая, каким образом исходные переменные (строки таблицы) были преобразованы

в серию новых переменных (столбцы таблицы). Так, к примеру, для переменной условие, имеющей 3 уровня, были сформированы 2 дополнительные переменные (1 и 2), представляющие собой серию контрастов между тремя условиями.

**Categorical Variables Codings**

| | | Frequency | Parameter coding | |
|---|---|---|---|---|
| | | | (1) | (2) |
| УСЛОВИЯ | на работе | 12 | 1.000 | .000 |
| | на улице | 17 | .000 | 1.000 |
| | в лесу | 17 | -1.000 | -1.000 |

**Рис. 24.4.** Кодирование категориальных переменных

## Последовательные оценки коэффициентов модели

Последовательные оценки коэффициентов модели показаны на рис. 24.5.

**Omnibus Tests of Model Coefficients**

| | | Chi-square | df | Sig. |
|---|---|---|---|---|
| Step 1 | Step | 15.920 | 2 | .000 |
| | Block | 15.920 | 2 | .000 |
| | Model | 15.920 | 2 | .000 |
| Step 2 | Step | 9.212 | 1 | .002 |
| | Block | 25.132 | 3 | .000 |
| | Model | 25.132 | 3 | .000 |

**Model Summary**

| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
|---|---|---|---|
| 1 | 47.501 | .293 | .391 |
| 2 | 38.289 | .421 | .563 |

**Рис. 24.5.** Последовательные оценки коэффициентов модели

Далее дана трактовка терминов, используемых профаммои в окне вывода и относящихся к последовательным оценкам коэффициентов модели.

- Step (Шаг) — методу Forward: LR (Прямой) понадобились всего 2 шага, чтобы включить все нужные переменные в уравнение регрессии. На самом деле позже в модель вводятся и другие переменные, однако на данном этапе будем считать, что их всего две.
- Chi-square (Хи-квадрат) — хи-квадрат шага (Step), блока (Block) или модели (Model). Это критерии статистической значимости воздействия на зависимую переменную всех предикторов заданной модели, блока или шага. На шаге 1 все три критерия $x^2$ оказались равными: для модели и шага потому, что на шаге 1 они тождественны, а для блока и модели потому, что модель содержит единственный блок. Большие значения критерия $\chi^2$ говорят о том, что

первая включенная переменная оказывает существенное влияние на зависимую переменную. Критерии $\chi^2$ шага 2 свидетельствуют о том, что включение второй переменной в модель значительно улучшает последнюю; обратите внимание, что критерий $\chi^2$ модели равен сумме критериев $\chi^2$ шага и модели на предыдущем шаге.

- -2 Log Likehood — эта величина характеризует модель и показывает, насколько хорошо она соответствует исходным данным. Чем меньше это значение, тем адекватнее сформированная модель. ,
- Cox & Snell R Square (R-квадрат Кокса и Снелла), Nagelkerke R Square (R-квадрат Нейджелкерка) — приближения значения $R^2$, показывающие долю влияния всех предикторов модели на дисперсию зависимой переменной.

## Классификационная таблица

В классификационной таблице сравниваются прогнозируемые значения зависимой переменной, рассчитанные по уравнению регрессии, и фактические наблюдаемые значения (рис. 24.6). При определении прогнозируемой величины SPSS вычисляет вероятность для каждого объекта и на основании этой вероятности присваивает объекту одно из двух значений дихотомической переменной. Если вероятность оказалась менее 0,5, то помощь оценивается как бесполезная (значение неременной помощь равно 0), в противном случае — как полезная (значение переменной помощь равно 1). Как показывают данные крайнего правого столбца таблицы, для 78,3 % объектов результаты прогноза оказались верными.

**Classification Table**[a]

| | Observed | | Predicted ПОМОЩЬ нет | Predicted ПОМОЩЬ есть | Percentage Correct |
|---|---|---|---|---|---|
| Step 1 | ПОМОЩЬ | нет | 13 | 8 | 61.9 |
| | | есть | 4 | 21 | 84.0 |
| | Overall Percentage | | | | 73.9 |
| Step 2 | ПОМОЩЬ | нет | 17 | 4 | 81.0 |
| | | есть | 6 | 19 | 76.0 |
| | Overall Percentage | | | | 78.3 |

a. The cut value is .500

**Рис. 24.6.** Классификационная таблица

## Переменные в уравнении

Таблица, представленная на рис. 24.7, демонстрирует эффекты включения переменных в уравнение на каждом шаге его построения. Строка Constant для каждого шага соответствует константе $B_0$ регрессионного уравнения.

**Variables in the Equation**

| | | B | S.E. | Wald | df | Sig. | Exp(B) |
|---|---|---|---|---|---|---|---|
| Step 1 [a] | УСЛОВИЯ | | | 11.413 | 2 | .003 | |
| | УСЛОВИЯ(1) | -.279 | .497 | .314 | 1 | .575 | .757 |
| | УСЛОВИЯ(2) | -1.457 | .495 | 8.659 | 1 | .003 | .233 |
| | Constant | .279 | .369 | .570 | 1 | .450 | 1.321 |
| Step 2b | АГРЕССИЯ | .652 | .270 | 5.827 | 1 | .016 | 1.920 |
| | УСЛОВИЯ | | | 9.842 | 2 | .007 | |
| | УСЛОВИЯ(1) | -.447 | .559 | .639 | 1 | .424 | .639 |
| | УСЛОВИЯ(2) | -1.658 | .611 | 7.371 | 1 | .007 | .190 |
| | Constant | -6.210 | 2.735 | 5.156 | 1 | .023 | .002 |

a. Variable(s) entered on step 1: УСЛОВИЯ.

b. Variable(s) entered on step 2: АГРЕССИЯ.

**Рис.** 24.7. Переменные в уравнении регрессии

Ниже дана трактовка остальных терминов, используемых программой в окне вывода и относящихся к переменным в уравнении регрессии.

- В — коэффициенты регрессионного уравнения, отражающие влияние соответствующих предикторов на зависимую переменную. Так, переменная агрессия оказывает положительное влияние на вероятность оказания помощи.
- S.E. (Стандартная ошибка) — мера изменчивости коэффициентов *B*.
- Wald (Критерий Вальда) — критерий значимости коэффициента *B* для соответствующего предиктора. Чем выше его значение (вместе с числом степеней свободы), тем выше значимость.
- Sig. (Значимость)— значимость по критерию Вальда.
- Exp(B) — величина ($e^B$), которая может использоваться для интерпретации результатов анализа наравне с коэффициентом *B* (вспомните о двух формах регрессионного уравнения, в одной из которых используются коэффициенты *B*, а *в* другой — $e^B$).

## Корреляционная матрица

На рис. 24.8 представлена обычная корреляционная матрица для всех переменных, вошедших в уравнение регрессии. Эта матрица включается в выводимые данные при установке флажка Correlations of estimates (Корреляции оценок). В случае значительных корреляций между предикторами нарушается условие их независимости, что может означать некорректность анализа.

## Параметры модели при исключении переменных

Все переменные модели проверяются на предмет их возможного исключения, и результаты такой проверки включаются в таблицу (рис. 28.9). Уровень значимости для каждой переменной сравнивается со значимостью 0,1, указанной в поле Removal (Вывод) диалогового окна Logistic Regression: Options (Логистическая

регрессия: Параметры). Если значимость переменной оказывается больше 0,1, происходит исключение переменной.

**Correlation Matrix**

| | | Constant | УСЛОВИЯ(1) | УСЛОВИЯ(2) | АГРЕССИЯ |
|---|---|---|---|---|---|
| Step 1 | Constant | 1.000 | -.137 | -.150 | |
| | УСЛОВИЯ(1) | -.137 | 1.000 | -.340 | |
| | УСЛОВИЯ(2) | -.150 | -.340 | 1.000 | |
| Step 2 | Constant | 1.000 | .151 | .375 | -.989 |
| | АГРЕССИЯ | -.989 | -.175 | -.392 | 1.000 |
| | УСЛОВИЯ(1) | .151 | 1.000 | -.264 | -.175 |
| | УСЛОВИЯ(2) | .375 | -.264 | 1.000 | -.393 |

**Рис. 24.8.** Корреляционная матрица

**Model if Term Removed**

| Variable | | Model Log Likelihood | Change in -2 Log Likelihood | df | Sig. of the Change |
|---|---|---|---|---|---|
| Step 1 | УСЛОВИЯ | -31.711 | 15.920 | 2 | .000 |
| Step 2 | АГРЕССИЯ | -23.750 | 9.212 | 1 | .002 |
| | УСЛОВИЯ | -27.048 | 15.808 | 2 | .000 |

**Рис. 24.9.** Результаты проверки переменных на возможность исключения из модели

## Не включенные в уравнение переменные

На рис. 24.10 приведены переменные, не включенные в уравнение регрессии. Столбец Sig. (Значимость) характеризует влияние данного предиктора на зависимую переменную без учета влияния остальных предикторов. Например, можно заключить, что переменная польза не оказывает значимого влияния на переменную помощь.

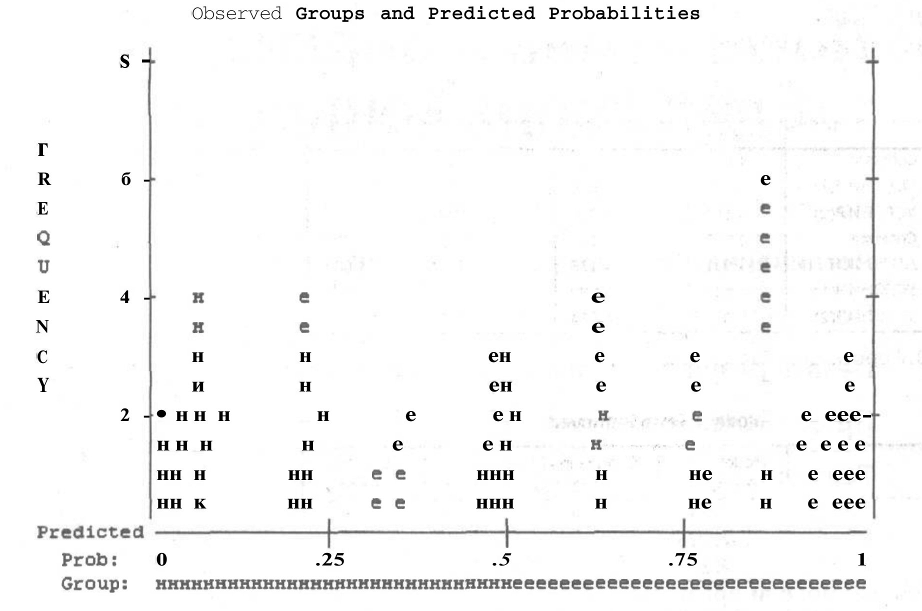
**Variables not in the Equation**

| | | | Score | df | Sig. |
|---|---|---|---|---|---|
| Step 1 | Variables | СИМПАТИЯ | 5.471 | 1 | .019 |
| | | АГРЕССИЯ | 7.604 | 1 | .006 |
| | | ПОЛЬЗА | 3.499 | 1 | .061 |
| | Overall Statistics | | 12.005 | 3 | .007 |
| Step 2 | Variables | СИМПАТИЯ | 3.622 | 1 | .057 |
| | | ПОЛЬЗА | 2.467 | 1 | .116 |
| | Overall Statistics | | 4.669 | 2 | .097 |

**Рис. 24.10.** Переменные, не включенные в уравнение регрессии

## Фактическая группировка и прогнозируемые вероятности

Диаграмма, показанная на рис. 24.11, генерируется в случае, если в окне Logistic Regression: Options (Логистическая регрессия: Параметры) установлен флажок Classification Plots (Классификационная диаграмма).




**Predicted Probability** is of Membership **for** есть
**The Cut Value is** .50
Symbols: и - нет
е - есть
**Each** Symbol Represents .5 Cases.

**Рис. 24.11.** Классификационная диаграмма

В диаграмме используются первые буквы градаций зависимой переменной: е (есть — помощь оказана) и н (нет — помощь не оказана). По горизонтальной оси отложены значения прогнозируемой вероятности, вычисляемые по уравнению регрессии, а по вертикальной оси — частоты. Таким образом, каждый столбик на диаграмме соответствует определенной предсказанной вероятности, а его высота — количеству объектов, для которых предсказана данная вероятность. В случае идеальной логистической регрессии все буквы н окажутся левее букв е, а разделять их будет вероятность 0,5. Как видно из диаграммы, некоторые столбики включают в себя обе буквы, что свидетельствует об ошибках предсказания (высота в два символа соответствует одному объекту). Символам н в правой части диаграммы и символам е в левой части диаграммы соответствуют неправильные предсказания относительно оказания помощи. О количестве правильных и неправильных предсказаний позволяет судить классификационная таблица (см. рис. 24.6).

# 25 Логлинейный анализ таблиц сопряженности



Логлинейные модели применяются для анализа таблиц сопряженности нескольких категориальных признаков[1]. Если анализируется сопряженность двух таких признаков (двухмерная таблица сопряженности), то вполне достаточно применение критерия $\chi^2$ (см. главу 8). Однако зачастую данные содержат существенно большее число категориальных признаков, и тогда визуальный анализ таблиц сопряженности становится невозможным. Например, визуально интерпретировать таблицу сопряженности полхклассхвузххобби практически невозможно. В подобных ситуациях на помощь исследователям приходят логлинейные модели, во многом сходные с регрессионными моделями и дисперсионным анализом.

В этой главе для иллюстрации иерархического логлинейного анализа мы используем файл helpLLM.sav. Этот файл содержит несколько модифицированные переменные файла helpLR.sav: переменная агрессия представлена в бинарном виде (1 — низкая, 2 — высокая), переменная симпатия имеет три категории (1 — слабая, 2 — средняя, 3 — сильная). Переменные помощь (2 градации) и условие (3 градации) оставлены без изменения. Таким образом, все переменные файла helpLLM.sav являются категориальными. В этой главе будет проводиться логлинейный анализ таблицы сопряженности агрессия×симпатия×условия×помощь.

## Понятие логлинейной модели

Логлинейная модель, в сущности, представляет собой множественную регрессионную модель, в которой категориальные переменные и их взаимодействия вы-

[1] Реализация логлинейного анализа требует полной установки программы SPSS, включая модули Advanced Models (Дополнительные модели) и Regression Models (Регрессионные модели).

ступают в качестве предикторов, а роль зависимой нерeменной играет натуральный логарифм частот категорий. Именно использование логарифмической меры обусловливает линейность модели. Для взаимодействия переменных агрессия, симпатия и условия уравнение будет иметь вид:

$$\ln(\text{частота}) = ц + \lambda А + КС + \lambda У + \lambda А{\times}С + \lambda А{\times}У + \lambda С{\times}У + \lambda А{\times}С{\times}У.$$

В этом уравнении частота — это частота текущей ячейки частотной таблицы, $\mu$ — общее среднее воздействия, эквивалентное константе во множественном регрессионном анализе, а каждое значение $\lambda$ представляет собой воздействие со стороны одной или более независимых переменных. Так, $\lambda$А является воздействием переменной агрессия, $\lambda$У и *КС* — соответственно переменных условия и симпатия, $\lambda$А×С — взаимодействия переменных агрессия и симпатия, $\lambda$А×С×У — взаимодействия переменных агрессия, симпатия и условия. Представленная здесь модель называется насыщенной, поскольку содержит все предикторы и их возможные взаимодействия. Однако насыщенная модель обычно не является оптимальной, так как редко все главные эффекты и взаимодействия оказываются значимыми. Как правило, существуют более предпочтительные альтернативы в виде ненасыщенных моделей, которые отражают лишь статистически значимые главные эффекты и взаимодействия переменных.

Подменю Loglinear (Логлинейный анализ) содержит три команды.

- General (Общая модель) — эта команда допускает вхождение в модель любых факторов и их взаимодействий и предполагает, что исследователь перед проведением анализа уже имеет гипотезы о составе модели.
- Logit (Логит-модель) — применение этой команды позволяет рассматривать дихотомические переменные как зависимые, а одну (или более) категориальную переменную как независимую. При этом зависимая дихотомическая переменная используется не для прогнозирования частот категорий, а для разделения всех категорий на две группы. Смысл понятия «логит» раскрывается в главе 24.
- Model selection (Подбор модели) — эта команда позволяет из всех возможных ненасыщенных моделей подобрать ту, которая в наибольшей степени соответствует исходным данным. Подбор осуществляется, как правило, автоматически. В результате выявляется совокупность значимых связей между категориальными переменными и вычисляются параметры ц и $\lambda$ логлинейной модели.

В данной главе рассматривается последний из перечисленных методов, который основан на применении иерархической логлинейной модели.

# Логлинейный метод подбора модели

Теоретически из насыщенной модели можно удалить любые элементы, получив, таким образом, произвольную ненасыщенную модель. Далее можно проверить состоятельность этой модели и в случае ее несоответствия исходным данным

перейти к анализу другой ненасыщенной модели. Однако большинство исследователей отдают предпочтение иерархическим логлинейным моделям, которые позволяют упорядочить процесс подбора окончательной состоятельной модели. Основной особенностью иерархических моделей является то, что присутствие какого-либо взаимодействия переменных означает присутствие всех взаимодействий, имеющих более низкий порядок, и главных эффектов этих переменных. Например, если в модели присутствует взаимодействие агрессия×симпатия, то в ней присутствуют главные эффекты переменных агрессия и симпатия; если в модели присутствует взаимодействие агрессия×симпатия×условия, то в ней также присутствуют взаимодействия агрессияхсимпатия, агрессия×условия и симпатия×условия, и т. д.

Существует три вспомогательных метода, которые предназначены для подбора адекватной модели. Все три метода оказываются весьма полезными и приводят к сходным результатам. Тем не менее окончательный выбор модели определяется не только статистическими результатами применения метода, но и пониманием исследователем особенностей исходных данных. Краткие описания методов приведены ниже, более подробные — в разделе «Представление результатов».

- Метод *исследования оценок параметров* предназначен для вычисления оценок параметров для насыщенной модели. Помимо обычных оценок параметров SPSS вычисляет стандартизованные оценки. Если значения последних невелики, то они не оказывают значимого влияния на модель и обычно исключаются.
- Метод *вычисления частичного критерия хи-квадрат* в дополнение к оценкам параметров модели SPSS вычисляет критерий $\chi^2$, характеризующий степень соответствия модели исходным данным. При помощи этого критерия проверяется, являются ли все однофакторные эффекты, а также эффекты более высоких порядков статистически значимыми (анализируются комбинации всех эффектов первого, второго и т. д. порядков). Обратите внимание на то, что отсутствие общей значимости эффектов второго порядка вовсе не означает, что все эффекты первого порядка не являются значимыми. Аналогично, из отсутствия общей значимости эффектов любого порядка не следует отсутствие значимости отдельных взаимодействий этого порядка. Вследствие этих двух особенностей в SPSS предусмотрена возможность раздельной проверки главных эффектов и эффектов взаимодействий.
- Суть метода *пошагового исключения,* заключающаяся в автоматической «подгонке» модели, схожа с методом исключения предикторов из уравнения регрессии: из насыщенной модели постепенно исключаются те элементы (переменные и их взаимодействия), которые не оказывают значимого воздействия. Данный метод построения модели относится к иерархическому логлинейному моделированию. Так, если обнаружено статистически значимое взаимодействие четырех переменных, то не проверяется (на предмет исключения из модели) взаимодействие трех из этих переменных, иначе модель не являлась бы иерархической по определению. Окончательный результат «подгонки» модели наиболее приемлем, если все оставшиеся в модели элементы оказываются статистически достоверными.

Иерархические логлинейные модели относятся к очень сложному разделу статистики, поэтому мы рекомендуем перед проведением логлинейного анализа обратиться к дополнительным источникам.

# Пошаговые алгоритмы вычислений

До применения логлинейного анализа для подбора модели нужно выполнить три подготовительных шага.

**Шаг 1** Создайте новый файл данных или подготовьте существующий. Все необходимые действия описаны в главе 3.

**Шаг 2** Запустите программу SPSS при помощи значка на рабочем столе или команды Пуск ▸ Программы ▸ SPSS for Windows ▸ SPSS 11.5 for Windows (Start ▸ Programs ▸ SPSS for Windows ▸ SPSS 11.5 for Windows) главного меню Windows. В открывшемся после запуска программы диалоговом окне SPSS for Windows щелкните на кнопке Cancel (Отмена).

После выполнения этого шага на экране появится окно редактора данных SPSS.

**Шаг 3** Откройте файл данных, с которым вы намерены работать (в нашем случае это файл helpLLM.sav). Если он расположен в текущей папке, то выполните следующие действия:

1. Выберите в меню File (Файл) команду Open ▸ Data (Открытие ▸ Данные) или щелкните на кнопке Open File (Открытие файла) панели инструментов.
2. В открывшемся диалоговом окне дважды щелкните на имени helpLLM.sav или введите его с клавиатуры и щелкните на кнопке OK.

Независимо от того, открыта программа SPSS только что или какие-то процедуры уже выполнялись, в верхней части главного окна должна присутствовать строка меню (она показана ниже). Пока строка меню присутствует на экране, доступны все команды анализа данных. При этом окно с данными видеть не обязательно.

File Edit View Data Transform Analyze Graphs Utilities Window Help

При работе со сводными таблицами или при редактировании диаграмм некоторые пункты строки меню могут исчезать или меняться. Чтобы вернуться к главному окну и стандартной строке меню, щелкните на кнопке свертывания или восстановления текущего окна.

После завершения шага 3 на экране должно присутствовать окно редактора данных.

**Шаг 4** В меню Analyze (Анализ) выберите команду Loglinear ▸ Model Selection (Логлинейный анализ ▸ Подбор модели). На экране появится диалоговое окно Model Selection Loglinear Analysis (Логлинейный анализ для подбора модели), показанное на рис. 25.1.

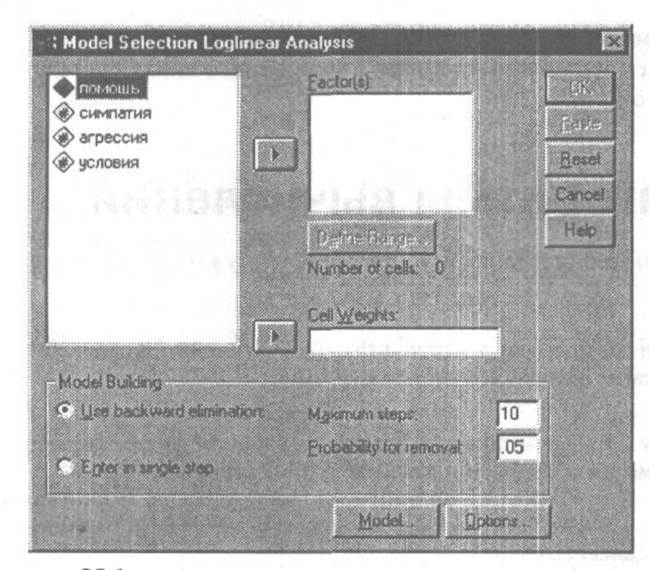



Рис. 25.1. Диалоговое окно Model Selection Loglinear Analysis

В этом диалоговом окне вы можете задать переменные для анализа, а также выбрать метод построения модели. Слева находится список всех переменных файла данных; чтобы включить переменную в анализ, достаточно выделить ее и переместить в список Factor(s) (Факторы), щелкнув на верхней кнопке со стрелкой. Данный метод применим только к категориальным переменным, поэтому вам придется пользоваться кнопкой Define Range (Определить диапазон). Эта кнопка открывает окно Loglinear Analysis: Define Range (Логлинейный анализ: Определение диапазона), показанное на рис. 25.2. В нем необходимо задать минимальный и максимальный уровни каждой переменной в полях Minimum (Минимум) и Maximum (Максимум).

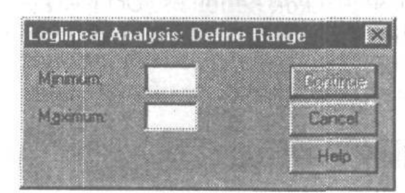



Рис. 25.2. Диалоговое окно Loglinear Analysis: Define Range

Поле Cell Weights (Ширина ячеек) диалогового окна Model Selection Loglinear Analysis (Логлинейный анализ для подбора модели) используется в том случае, если ваша модель содержит структурные нули. Понятие структурных нулей в этой книге не рассматривается; при необходимости обратитесь к руководству пользователя SPSS.

Диалоговые окна задания переменных для команд Loglinear ▸ General (Логлинейный анализ ▸ Общий анализ) и Loglinear ▸ Logit (Логлинейный анализ ▸ Логит-модель) отличаются тем, что позволяют задавать в качестве факторов не только категориальные, но и количественные переменные (ковариаты). Помимо этого при использовании

логит-моделей необходимо задание одной или нескольких бинарных зависимых переменных, каждая из которых используется в анализе для деления независимых переменных на две категории.

В группе Model Building (Построение модели), расположенной в нижней части диалогового окна Model Selection Loglinear Analysis (Логлинейный анализ для подбора модели), имеется два переключателя. Переключатель Use Backward Elimination (Использовать пошаговое исключение) определяет построение модели методом пошагового исключения: из насыщенной модели в результате серии шагов исключаются все элементы, не оказывающие значимого влияния на модель. Если вам требуется модель со всеми элементами, установите переключатель Enter in single step (Ввести все элементы). Если вы решили задействовать метод пошагового исключения, в поле Probability for removal (Значимость для исключения) укажите величину значимости, которая будет служить критерием исключения элементов из модели. По умолчанию используется значение $p = 0,05$, и нередко оно устраивает исследователей. Под *шагом* понимается удаление очередного элемента из модели; в поле Maximum Steps (Максимум шагов) вы можете задать максимальное количество шагов для того, чтобы процедура «преждевременно» прекратила процесс исключения предикторов. По умолчанию максимальное число шагов равно 10.

После того как метод построения модели и переменные, входящие в ее состав, заданы, можно переходить к определению других параметров команды, используя кнопки Options (Параметры) и Model (Модель).

При щелчке на кнопке Options (Параметры) открывается диалоговое окно Loglinear Analysis: Options (Логлинейный анализ: Параметры), представленное на рис. 25.3. Это окно позволяет настроить параметры генерируемых командой итоговых данных, а также задать некоторые технические аспекты ее выполнения (последние в этой книге не рассматриваются). Флажок Frequencies (Частоты) включает в выводимые данные таблицу частот модели, а флажок Residuals (Остатки) – разности между фактическими и прогнозируемыми частотами, рассчитываемыми при помощи модели.

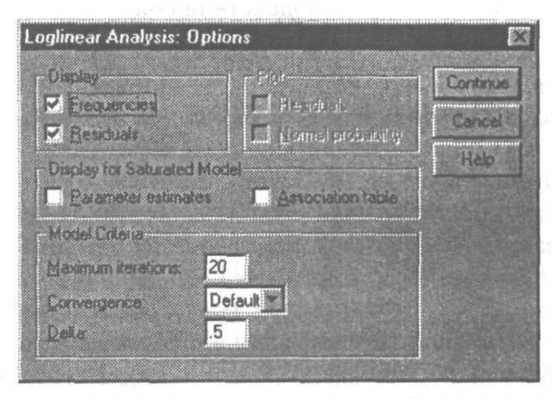



**Рис. 25.3.** Диалоговое окно Loglinear Analysis: Options

Нередко используются флажки Parameter estimates (Оценки параметров) и Assosiation table (Таблица ассоциаций). Для насыщенной модели эти флажки позволяют включить в выводимые данные оценки параметров, что может быть полезно при выборе модели, и таблицу ассоциаций, содержащую критерии $\chi^2$ для всех главных эффектов и взаимодействий. Оба флажка при исследовании насыщенной модели рекомендуется оставлять установленными. Если же для анализа необходима ненасыщенная модель, включающая лишь некоторые взаимодействия и факторы, то, щелкнув на кнопке Model (Модель), следует открыть диалоговое окно Loglinear Analysis: Model (Логлинейный анализ: Модель), представленное на рис. 25.4.

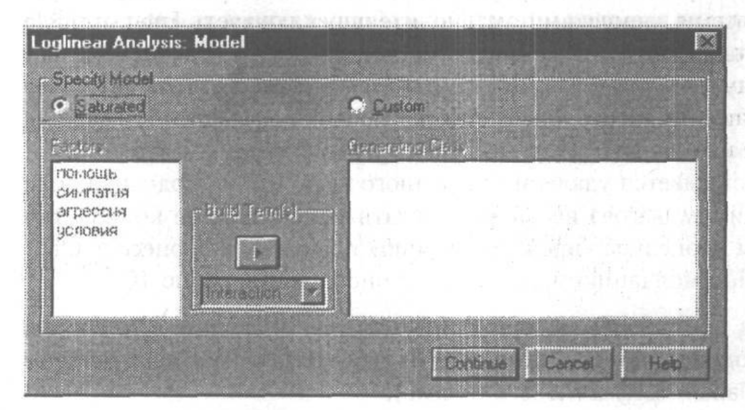



**Рис. 25.4.** Диалоговое окно Loglinear Analysis: Model

Если в группе Specify Model (Задать модель) установить переключатель Custom (Настройка), можно щелчками выбирать факторы в списке Factors (Факторы). Раскрывающийся список в центре окна позволяет выбрать один из типов взаимодействий выбранных факторов: Main effects (Главные эффекты), Interactions (Взаимодействия), All 2-Way (Все 2-факторные), АН 3-Way (Все 3-факторные), АН 4-Way (Все 4-факторные) и All 5-Way (Все 5-факторные). Поскольку команда Loglinear ▸ Model Selection (Логлинейный анализ ▸ Подбор модели) строит иерархические модели, то при выборе взаимодействий все взаимодействия более низких порядков включаются в модель автоматически.

Задание модели вручную требуется в тех случаях, когда есть предварительные гипотезы о составе ненасыщенной модели. Иногда модель задается для того, чтобы ограничить количество рассматриваемых взаимодействий, если взаимодействия высоких порядков вызывают трудности в интерпретации. Если вы хотите настроить модель вручную, установите переключатель Custom (Настройка), тогда вы сможете включить в модель те главные факторы и взаимодействия, которые вам необходимы. Главные факторы включаются в модель следующим образом: имя соответствующего фактора выделяется в списке Factors (Факторы) и с помощью кнопки с направленной вправо стрелкой перемещается в правый список. Если требуется включить в модель взаимодействие факторов, то следует щелчками выделить в списке Factors & Covariates (Факторы и ковариаты) имена всех

факторов, участвующих во взаимодействии, затем в раскрывающемся списке Build Term(s) (Создание элементов) выбрать пункт Interaction (Взаимодействие) и, наконец, как и в предыдущем случае, щелкнуть на кнопке с направленной вправо стрелкой. Чтобы построить сразу несколько взаимодействий определенного порядка для выбранных факторов, нужно выполнить те же действия, что и для одиночного взаимодействия, но в раскрывающемся списке выбрать пункт АИ 2-Way (Все 2-факторные), АН 3-Way (Все 3-факторные), All 4-Way (Все 4-факторные) или All 5-Way (Все 5-факторные) в зависимости от желаемого результата.

В следующем примере проводится иерархический логлинейный анализ переменных агрессия, симпатия, условия и помощь, имеющих соответственно 2, 3, 3 и 2 уровня. Мы будем исследовать насыщенную модель, а в выводимые данные включим оценки параметров и критерии $\chi^2$ для отдельных эффектов. Затем мы применим метод пошагового исключения, чтобы удалить из насыщенной модели те элементы, которые не оказывают значимого воздействия.

**Шаг 5** После выполнения предыдущего шага у вас должно быть открыто диалоговое окно Model Selection Loglinear Analysis (Логлинейный анализ для подбора модели), показанное на рис. 25.1. Если вы уже успели поработать с этим окном, щелкните на кнопке Reset (Сброс).

1. Щелкните сначала на переменной помощь, чтобы выделить ее, а затем — на верхней кнопке со стрелкой, чтобы переместить переменную в список Factor(s) (Факторы).
2. Щелкните на кнопке Define Range (Определить диапазон), чтобы открыть диалоговое окно Loglinear Analysis: Define Range (Логлинейный анализ: Определение диапазона), показанное на рис. 25.2.
3. В поле Minimum (Минимум) введите число 0, нажмите клавишу Tab, чтобы переместить фокус ввода в поле Maximum (Максимум), введите число 1 и щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Model Selection Loglinear Analysis (Логлинейный анализ для подбора модели).
4. Повторите первые три действия для переменных симпатия, агрессия и условия, используя для задания значений переменных симпатия и условие числа 1 и 3, а для задания значений переменной агрессия числа 1 и 2.
5. Щелкните на кнопке Options (Параметры), чтобы открыть диалоговое окно Loglinear Analysis: Options (Логлинейный анализ: Параметры), показанное на рис. 25.3.
6. Установите флажки Parameter estimates (Оценки параметров) и Assotiation table (Таблица ассоциаций), а затем щелкните на кнопке Continue (Продолжить), чтобы вернуться в диалоговое окно Model Selection Loglinear Analysis (Логлинейный анализ для подбора модели).
7. Щелкните на кнопке OK, чтобы открыть окно вывода.

После выполнения шага 5 программа автоматически активизирует окно вывода. Для просмотра результатов вы при необходимости можете воспользоваться вертикальной и горизонтальной полосами прокрутки. Обратите внимание на

стандартную строку меню в верхней части окна вывода: ее присутствие позволяет выполнять любые статистические операции, не переключаясь обратно в окно редактора данных.

# Печать результатов и выход из программы

Ниже описана типичная процедура печати результатов статистического анализа (или нескольких анализов). После выполнения шага 5 должно быть открыто окно вывода.

**Шаг 6** В окне вывода укажите фрагменты, выводимые на печать (см. раздел «Окно вывода» в главе 2), в меню File (Файл) выберите команду Print (Печать), при необходимости задайте параметры печати и щелкните на кнопке OK.

Последнее, что необходимо сделать после завершения исследования и печати результатов, — это выйти из программы SPSS.

**Шаг 7** Для выхода из программы в меню File (Файл) выберите команду Exit (Выход).

Иногда после выполнения команды Exit (Выход) на экране могут появляться небольшие диалоговые окна с вопросом о необходимости сохранения сделанных в файлах изменений и кнопками, описывающими возможные варианты ответа. Для завершения работы просто щелкайте на соответствующих кнопках.

# Представление результатов

В этом разделе приводятся фрагменты данных, генерируемые программой при выполнении шага 5. Полностью эти данные занимают приблизительно 20 страниц.

Напоминаем, что если выводимые данные занимают слишком много места, они представлены в окне вывода лишь частично. В этом случае в том месте окна вывода, в котором скрыты дополнительные данные, присутствует значок в виде красного треугольника. Чтобы увидеть итоговые данные полностью, дважды щелкните на имеющемся фрагменте.

## Проверка на равенство нулю взаимодействий порядка *K* и выше

В таблице, показанной на рис. 25.5, представлены результаты проверки эффекта от взаимодействий указанного и выше порядка. Для проверки используется два варианта критерия $\chi^2$. Оба критерия оценивают, являются ли ожидаемые частоты в ячейках для соответствующей модели значимо отличающимися от наблюдаемых

частот или нет. Если отличие значимо, то гипотеза об отсутствии взаимодействия отвергается и делается вывод о связи соответствующих переменных. Так, первая строка соответствует 4-му порядку взаимодействия (значение в столбце K равно 4) и характеризует общий (суммарный) эффект от взаимодействий 4 и более переменных; высокий уровень значимости и низкое значение $x^2$ говорят о том, что взаимодействие всех 4 переменных не оказывает большого влияния на модель. То же касается взаимодействия 3-го порядка и выше. Противоположный вывод можно сделать об остальных взаимодействиях: $\chi^2$ возрастает, а значимость убывает с убыванием порядка. Последние две строки свидетельствуют о том, что среди всех взаимодействий имеются такие, которые оказывают на модель значимое влияние.

Tests that K-way and higher order effects are zero.

| K | DF | L.R. Chisq | Prob | Pearson Chisq | Prob | Iteration |
|---|---|---|---|---|---|---|
| 4 | 4 | .397 | .9827 | .212 | .9948 | 4 |
| 3 | 16 | 21.784 | .1503 | 17.649 | .3448 | 4 |
| 2 | 29 | 54.241 | .0030 | 47.940 | .0149 | 2 |
| 1 | 35 | 58.726 | .0072 | 52.609 | .0283 | 0 |

**Рис. 25.5.** Проверка эффекта от взаимодействий порядка K и выше

Ниже дана трактовка терминов, используемых программой в окне вывода.

- K — порядок взаимодействий (4 — порядок 4 и выше, 3 — порядок 3 и выше, и т. д.).
- DF (Число степеней свободы) — число степеней свободы для взаимодействий порядка *K* и выше.
- Pearson Chisq (Критерий хи-квадрат Пирсона) — критерий $\chi^2$ для определения вероятности того, что общий эффект от взаимодействий порядка *K* и выше равен нулю.
- L.R. Chisq (Критерий хи-квадрат правдоподобия) — критерий $\chi^2$ Для определения вероятности того, что общий эффект от взаимодействий порядка *K* и выше равен нулю (по утверждению его авторов этот критерий более «правдоподобный», чем критерий хи-квадрат Пирсона).
- Prob (Вероятность) — вероятность того, что общий эффект от взаимодействий порядка *K* и выше равен нулю. Чем меньше вероятность, тем большее влияние эффект оказывает на модель.
- Iter (Итерации) — число итераций, потребовавшееся для вычисления значений $\chi^2$.

# Проверка на равенство нулю взаимодействий только порядка *K*

Фрагмент, приведенный па рис. 25.6, аналогичен предыдущему, однако характеризует взаимодействия только указанного порядка. Как можно видеть, значимое воздействие на модель оказывают взаимодействия 2-го и 3-го порядков.

```
Tests that K-way effects are zero.

     K    DF   L.R. Chisq    Prob  Pearson Chisq    Prob  Iteration

     1     6       4.486    .6113          4.669   .5870          0
     2    13      32.456    .0021         30.291   .0043          0
     3    12      21.387    .0450         17.437   .1339          0
     4     4        .397    .9827           .212   .9948          0
```

**Рис. 25,6.** Проверка эффекта от взаимодействий только порядка K

Ниже дана трактовка терминов, используемых программой в окне вывода.

- K — порядок взаимодействий (1 — порядок 1, 2 — порядок 2, и т. д.).
- DF (Число степеней свободы) — число степеней свободы для взаимодействий указанного порядка.
- Pearson Chisq (Критерий хи-квадрат Пирсона) — значение, характеризующее вероятность того, что общий эффект от взаимодействий указанного порядка равен нулю.
- L.R. Chisq (Критерий хи-квадрат правдоподобия) — значение, характеризующее вероятность того, что общий эффект от взаимодействий указанного порядка равен нулю (по утверждению его авторов, этот критерий более «правдоподобный», чем критерий хи-квадрат Пирсона).
- Prob (Вероятность) — вероятность того, что общий эффект от взаимодействий указанного порядка равен нулю. Чем меньше вероятность, тем большее влияние эффект оказывает на модель.
- Iter (Итерации) — для данной таблицы число итераций всегда равно 0.

## Проверка частичных ассоциаций

На рис. 25.7 приведены критерии $x^2$ Для каждого из воздействий насыщенной модели. Воздействия с высоким значением $\chi^2$ оказывают значимое влияние на модель; чем ниже значение $\chi^2$, тем ниже значимость. В данном примере главные эффекты переменных не являются статистически значимыми, а к значимым можно отнести двухфакторные взаимодействия помощь×условие и симпатия×помощь, а также взаимодействие трех переменных помощь×симпатия×агрессия. Поскольку эти частичные критерии $\chi^2$ не обязательно являются независимыми, их доля в итоговом значении $\chi^2$ для насыщенной модели может отличаться от их доли в ненасыщенной модели.

Далее дана трактовка терминов, используемых программой в окне вывода.

- DF (Число степеней свободы) — число степеней свободы соответствующего эффекта.
- Partial Chisq (Частичный критерий хи-квадрат) — критерий $\chi^2$ для отдельного эффекта.

- Prob (Вероятность) — вероятность того, что величина эффекта равна нулю. Чем меньше вероятность, тем большее влияние эффект оказывает на модель.
- Iter (Итерации) — число итераций, которые потребовались для вычисления соответствующей частичной ассоциации.

Teats of PARTIAL associations.

| Effect Name | DF | Partial Chisq | Prob | Iter |
|---|---|---|---|---|
| ПОМОЩЬ*СИМПАТИЯ*АГРЕССИЯ | 2 | 6.288 | .0431 | 3 |
| ПОМОЩЬ*СИМПАТИЯ*УСЛОВИЯ | 4 | .091 | .9990 | 5 |
| ПОМОЩЬ*АГРЕССИЯ*УСЛОВИЯ | 2 | .545 | .7615 | 5 |
| СИМПАТИЯ*АГРЕССИЯ*УСЛОВИЯ | 4 | 2.441 | .6552 | 6 |
| ПОМОЩЬ*СИМПАТИЯ | 2 | 7.430 | .0244 | 3 |
| ПОМОЩЬ*АГРЕССИЯ | 1 | 1.956 | .1620 | 4 |
| СИМПАТИЯ*АГРЕССИЯ | 2 | 1.944 | .3784 | 4 |
| ПОМОЩЬ*УСЛОВИЯ | 2 | 16.593 | .0002 | 3 |
| СИМПАТИЯ*УСЛОВИЯ | 4 | 2.277 | .6850 | 3 |
| АГРЕССИЯ*УСЛОВИЯ | 2 | 1.230 | .5406 | 4 |
| ПОМОЩЬ | 1 | .348 | .5551 | 2 |
| СИМПАТИЯ | 2 | 1.605 | .4481 | 2 |
| АГРЕССИЯ | 1 | 1.398 | .2370 | 2 |
| УСЛОВИЯ | 2 | 1.134 | .5673 | 2 |

**Рис. 25.7.** Проверка частичных ассоциаций

## Оценки параметров **модели**

Фрагмент результатов, представленный на рис. 25.8, содержит оценки параметров, представляющие собой значения $\lambda$ из логлинейного уравнения: для трехфакторного взаимодействия помощь×симпатия×агрессия и двухфакторного взаимодействия помощь×условия.

Estimates lor Parameters.

.

.

ПОМОЩЬ*СИМПАТИЯ*АГРЕССИЯ

| Parameter | Coeff. | Std. Err. | Z-Value | Lower 95 CI | Upper 95 CI |
|---|---|---|---|---|---|
| 1 | -.2855257838 | .25774 | -1.10781 | -.79070 | .21964 |
| 2 | .4433451616 | .23575 | 1.88058 | -.01872 | .90541 |

.

.

.

ПОМОЩЬ*УСЛОВИЯ

| Parameter | Coeff. | Std. Err. | Z-Value | Lover 95 CI | Upper 95 CI |
|---|---|---|---|---|---|
| 1 | .0368824557 | .24631 | .14974 | -.44588 | .51965 |
| 2 | .4833207809 | .23860 | 2.02565 | .01566 | .95098 |

**Рис. 25.8.** Оценки параметров модели

Поскольку сумма всех параметров должна быть равна нулю, на основе полученных данных можно вычислить другие. Так, в нижней строке и правом столбце табл. 25.1 показаны значения *K*, вычисленные на основе представленных данных для взаимодействия помощьхусловия, исходя из того, что сумма значений в строках и столбцах должна быть равна 0.

**Таблица 25.1.** Пример кодирования взаимодействия

| Уровни переменной «помощь» | Уровни переменной «симпатия» | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| 0 | 0,0369 | 0,4833 | -0,4798 |
| **1** | **-0,0369** | -0,4833 | **0,4798** |

Полученные оценки можно использовать для интерпретации распределения частот. Например, из таблицы (нижняя строка) видно, что при наибольшем уровне симпатии (3) помощь оказывается чаще. Но лучше для этих целей использовать таблицы сопряженности, воспользовавшись командой Descriptive Statistics ▸ Crosstabs (Описательные статистики ▸ Таблицы сопряженности).

Ниже дана трактовка терминов, используемых программой в окне вывода и относящихся к оценкам параметров модели.

- Parameter (Параметр) — номер параметра. Нумерация параметров производится от наименьших значений независимых переменных к наибольшим (в качестве примера кодирования взаимодействий см. табл. 25.1).
- Coef. (Коэффициент) — значение $\lambda$ в логлинейном уравнении.
- Std. Err. (Стандартная ошибка) — мера изменчивости коэффициента.
- Z-Value (z-значение) — *z*-значение коэффициента. Значения больше 1,96 (меньше -1,96) статистически достоверны (имеют уровень значимости меньше 0,05).
- Lower 95 CI (Нижняя граница доверительного интервала в 95 %), Upper 95 CI (Верхняя граница доверительного интервала в 95 %) — границы интервала, в который фактическое (не приближенное) значение коэффициента попадает с вероятностью *95 %*.

## Обратное пошаговое исключение

Результаты обратного пошагового исключения являются, пожалуй, наиболее важными среди всех остальных, поскольку дают представление о наиболее состоятельной модели (рис. 25.9).

В представленном фрагменте результатов содержатся критерии $\chi^2$, их значимости и все взаимодействия, участвующие в процессе пошагового исключения (шаги 2-5 для экономии места опущены).

```
Backward Elimination (p - .050) for DESIGN 1 with generating class
     .
     .
     .
If Deleted Simple Effect is                    DF   L.R. Chisq Change    Prob  Iter
 ПОМОЩЬ*СИМПАТИЯ*АГРЕССИЯ*УСЛОВИЯ              4                .397    .9827     4

Step 1
  The best model has generating class
      ПОМОЩЬ*СИМПАТИЯ*АГРЕССИЯ
      ПОМОЩЬ * СИНПАТИЯ* УС ЛОВИЯ
      ПОМОЩЬ*АГРЕССИЯ*УСЛОВИЯ
      СИМПАТИЯ*АГРЕССИЯ*УСЛОВИЯ

  Likelihood ratio chi square -     .39680    DF • 4   P -  .983
     .
     .
     .
Step 6
  The best model has generating class
      ПОМОЩЬ*СИМПАТИЯ*АГРЕССИЯ
      ПОНОЩЬ *УСЛОВИЯ
  Likelihood ratio chi square -    13.26806    DF » 20   P •  .866

If Deleted Simple Effect is                    DF   L.R. Chisq Change    Prob  Iter
 ПОМОЩЬ*СИМПАТИЯ*АГРЕССИЯ                      2              11.944    .0025     3
 ПОМОЩЬ*УСЛОВИЯ                                2              15.920    .0003     2

Step 7
  The best model has generating class
      ПОНОЩЬ *СИМПАТИЯ* АГРЕССИЯ
      ПОМОЩЬ*УСЛОВИЯ

  Likelihood ratio chi square -    13.26806    DF - 20   P -  .866

The final model has generating class
    ПОМОЩЬ*СИМПАТИЯ*АГРЕССИЯ
    ПОМОЩЬ*УСЛОВИЯ
```

**Рис. 25.9.** Некоторые результаты обратного пошагового исключения

Процесс пошагового исключения начинается с насыщенной модели. Сначала SPSS вычисляет значение $\chi^2$ для модели, включающей взаимодействия четвертого и более низких порядков; это значение равно 0, поскольку изначально модель является насыщенной. Далее устанавливается, что при исключении взаимодействия четвертого порядка изменение величины $\chi^2$ не является значимым, и исключение производится. Затем анализу подвергается модель, состоящая из взаимодействия третьего и более низких порядков; из модели исключаются те взаимодействия, чье влияние оказывается незначимым. Весь процесс занимает 7 шагов, и в результате формируется иерархическая модель, в которую входят взаимодействия помощь×симпатия×агрессия и помощь×условия. Оба взаимодействия статистически достоверны. Значение критерия $\chi^2$ для всей модели составляет 13,268, а число степеней свободы равно 20. Следует отметить, что критерий $\chi^2$ для всей модели оценивает расхождение между наблюдаемыми частотами в исходной таблице сопряженности и ожидаемыми частотами, вычисленными при

помощи модели. Поскольку результат оказался незначимым, созданная модель адекватна исходным данным: ожидаемые в соответствии с моделью частоты вполне точно соответствуют наблюдаемым частотам. Для содержательной интерпретации модели следует обратиться к таблицам сопряженности, соответствующим значимым компонентам модели, в данном случае к таблицам помощь×симпатияхагрессия и помощь×условия.

## Частоты и остатки

Во фрагменте, представленном на рис. 25.10 (для экономии места показаны только начало и конец фрагмента), для каждой ячейки исходной таблицы сопряженности генерируются фактические и прогнозируемые значения.

Observed, Expected Frequencies and Residuals.

| Factor | Code | OBS count | EXP count | Residual | Std Resid |
|---|---|---|---|---|---|
| ПОМОЩЬ | нет | | | | |
| СИМПАТИЯ | 1 | | | | |
| АГРЕССИЯ | 1 | | | | |
| УСЛОВИЯ | на работ | .0 | 1.4 | -1.43 | -1.20 |
| УСЛОВИЯ | на улице | 4.0 | 3.1 | .90 | .51 |
| УСЛОВИЯ | в лесу | 1.0 | .5 | .52 | .76 |
| АГРЕССИЯ | 2 | | | | |
| УСЛОВИЯ | на работ | 2.0 | 1.1 | .86 | .80 |
| УСЛОВИЯ | на улице | 2.0 | 2.5 | -.48 | -.30 |
| УСЛОВИЯ | в лесу | .0 | .4 | -.38 | -.62 |
| СИМПАТИЯ | 2 | | | | |
| АГРЕССИЯ | 1 | | | | |
| УСЛОВИЯ | на работ | 3.0 | 2.6 | .43 | .27 |

.
.
.

Goodness-of-fit test statistics

Likelihood ratio chi square - 13.26806 DF - 20 P - .866
Pearson chi square • 10.07265 DF - 20 P - .967

**Рис. 25.10.** Фактические и прогнозируемые значения для каждой ячейки исходной таблицы сопряженности

Следует отметить, что при анализе насыщенной модели программа добавляет 0,5 к каждой частоте (чтобы не было ячеек с нулевыми частотами). Прогнозируемые частоты конечной модели формируются по методу пошагового исключения. Величина остатка указывает на возможные несоответствия между моделью и данными. Поскольку величины $\chi^2$ невелики, а уровни значимости приближаются к 1, есть все основания полагать, что окончательная модель адекватно отражает данные.

# Терминология, используемая при выводе

Далее дана трактовка терминов, используемых программой в окне вывода и относящихся к частотам и остаткам.

- OBS Count (Наблюдаемое значение) — фактическое значение частоты, полученное на основе исходных данных.
- EXP Count (Ожидаемое значение) — значение частоты, рассчитанное на основе созданной модели.
- Residual (Остаток), Std Resid (Стандартизированный остаток) — разность между наблюдаемым и ожидаемым значениями; чем больше эта величина, тем меньше точность модели.
- Coodness-of-fit test statistics (Критерии согласия) — критерии $\chi^2$ для определения степени адекватности модели исходным данным. Чем выше эти значения и чем ниже соответствующие уровни значимости, тем хуже модель соответствует данным; в данном случае $p > 0,1$ и приближается к 1, следовательно, модель хорошо согласуется с данными.

# 26 Анализ остатков



Эта глава может показаться несколько необычной в сравнении с другими главами книги. В ней вы не найдете привычных инструкций по выполнению тех или иных статистических анализов; более того, здесь не рассматриваются новые команды SPSS. На этот раз предметом обсуждения для нас станут остатки — статистическое понятие, неоднократно встречавшееся в различных главах книги. Таблицы сопряженности, однофакторный и многофакторный типы дисперсионного анализа, регрессионный анализ, логлинейные модели — во всех этих разделах статистики используется понятие остатка. Анализ остатков является весьма непростым и требует определенной опытности в статистическом анализе. Анализ остатков выделен в отдельную главу для того, чтобы рассмотреть некоторые его аспекты, общие для всех статистических процедур. Если бы мы занимались анализом остатков в каждой из глав, где это было бы уместно, то объем книги был бы значительно больше. Кроме того, наличие подобного материала в первых главах книги могло бы запутать тех, кто не имеет большого опыта статистического анализа.

Итак, данная глава посвящена остаткам. После введения понятия остатков будут рассмотрены примеры, иллюстрирующие анализ остатков на примере линейной регрессии и общей логлинейной модели. Завершает главу описание различных методов анализа остатков в SPSS.

## Понятие остатков

Многие сложные статистические процедуры связаны с созданием теоретической модели.

- Анализ таблиц сопряженности с использованием критерия $x^2$ применяется для вычисления ожидаемых частот; эти ожидаемые частоты вычисляются на основе модели, включающей две категориальные переменные.
- Дисперсионный анализ предназначен для исследования влияний факторов и их взаимодействий. Здесь моделью является совокупность независимых переменных, которые оказывают влияние на одну или несколько зависимых

переменных. Именно эта общность позволяет объединить несколько рассмотренных ранее команд SPSS в подменю General Linear Model (Общая линейная модель).

- В регрессионном анализе одна (или более) переменная-предиктор позволяет прогнозировать значения зависимой переменной (критерия). Путем регрессионного анализа можно выяснить, существует ли значимое влияние каждого предиктора на критерий и оценить долю этого влияния. Таким образом, можно говорить о том, что в регрессионном анализе используется модель предикторов.
- Логлинейный анализ исследует отношения между категориальными переменными, создавая и улучшая модель, включающую факторы и взаимодействия между ними.

Как можно заметить, все модели включают исходные данные и прогнозируемые значения. Целью является создание наилучшей модели в том смысле, что такая модель, будучи максимально простой, должна с высокой точностью прогнозировать значения зависимой переменной (критерия). Так как модель — это всегда упрощение, между фактическим и прогнозируемым значениями существует разность, называемая остатком. Величина остатка позволяет судить о том, насколько точно модель описывает данные. Остатки также используются для выявления объектов, которые содержат «нетипичные» значения переменных и не соответствуют анализируемой модели.

Далее представлен пример, иллюстрирующий оценку модели, получаемую путем простого регрессионного анализа.

# Остатки и линейная регрессия

Вернемся к ситуации, которая была описана в главе 17. Исследователь выдвинул гипотезу о том, что студенты с низким показателем нервной возбудимости имеют низкие результаты на экзаменах, и наоборот, легковозбудимые студенты показывают высокие результаты (данные содержатся в файле exam.sav, описанном в главе 17). Таким образом, для проверки достоверности этой гипотезы нужно оценить близость указанной зависимости к линейной. В пошаговых процедурах (см. шаги 4 и 5) главы 17 была создана соответствующая модель. Кратко, эта модель представляет собой регрессионное уравнение:

$$\text{тест}_{\text{истина}} - \textit{константа} + \textit{коэффициент} \times \text{трев} + \textit{остаток},$$

или:

$$\text{тест}_{\text{истина}} = 9{,}3114 + 0{,}6751 \times \text{трев} + \textit{остаток}.$$

Для того чтобы оценить, насколько точно эта модель по заданным значениям переменной трев прогнозирует значения переменной тест, необходимо вычислить величины остатков для каждого из объектов файла (для которых значения этих

двух переменных известны). Остатки удобно представлять в графической форме, поэтому предусмотрены различные способы визуальной интерпретации остатков. Так, вы можете включить в выводимые данные гистограмму, диаграмму нормальных вероятностей, а также различные виды диаграмм для отображения остатков, прогнозируемых и фактических значений. Напомним (см. главу 17), что для управления параметрами диаграмм в диалоговом окне Linear Regression (Линейная регрессия) имеется кнопка Plots (Диаграммы).

На рис. 26.1 представлена гистограмма, по горизонтальной оси которой отображены стандартизованные остатки (остатки, деленные на оценку стандартной ошибки), а по вертикальной — количества объектов, соответствующие диапазонам стандартизованных остатков.

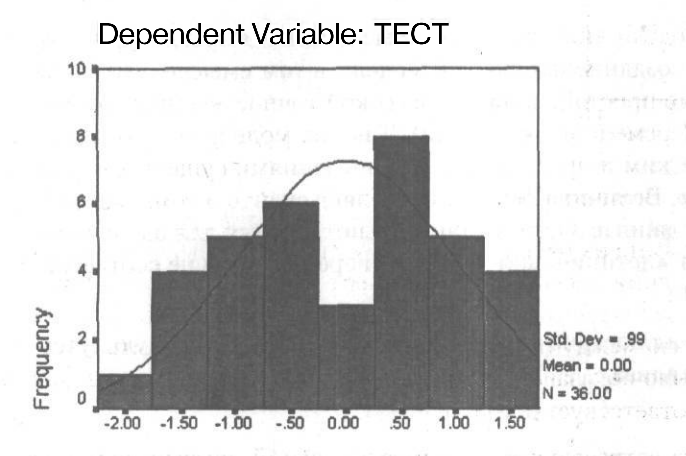



**Рис. 26.1.** Гистограмма остатков

Величина стандартного отклонения остатков на гистограмме равна 1 или близка к ней (из-за ошибки округления), а среднее значение равно 0. Это объясняется тем, что остатки стандартизованы. На гистограмму также нанесена нормальная кривая. Наибольший интерес в исследовании остатков представляет разность между распределением остатков и нормальной кривой. На приведенной гистограмме столбики почти соответствуют нормальной кривой, но в середине они находятся немного под ней, а по краям, особенно справа, они чуть выше кривой. Это может свидетельствовать о том, что, возможно, данные обнаруживают криволинейную зависимость.

Диаграмма нормальных вероятностей (накопленных частот), изображенная на рис. 26.2, является более сложной, однако нагляднее демонстрирует характер зависимости между переменными трев и тест. По горизонтальной оси отложены наблюдаемые накопленные вероятности, а по вертикальной — ожидаемые накопленные вероятности. В случае идеальной модели зависимость между наблюдаемыми и ожидаемыми накопленными вероятностями является строго линейной,

и все точки лежат на одной прямой. Любые отклонения точек от этой прямой соответствуют ошибкам модели.

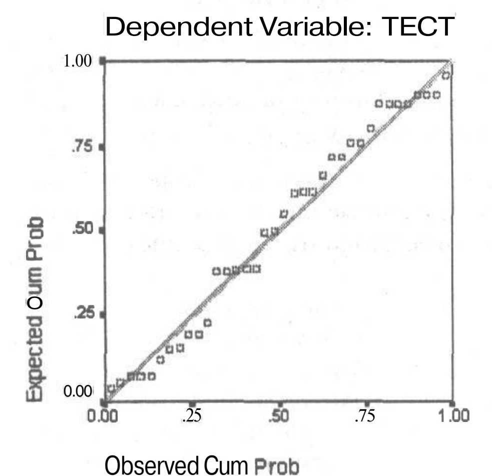



**Рис. 26.2.** Диаграмма нормальных вероятностей (накопленных частот)

Как можно судить по диаграмме из нашего примера, в диапазоне выше среднего (выше 0,5) ожидаемые значения несколько превышают наблюдаемые, а в нижнем диапазоне (менее 0,5) наблюдается обратная тенденция. Подобный результат также указывает на то, что между переменными трев и тест может быть нелинейная зависимость. Как было показано в главе 17, модель, включающая квадрат переменной трев, лучше соответствует данным.

Существуют и другие часто встречающиеся типы отклонений, указывающие па несоответствие уравнения регрессии фактическим данным. С помощью приведенных диаграмм вы всегда сможете предложить собственные гипотезы по устранению недостатков модели.

# Остатки и логлинейные модели

В главе 25 был рассмотрен пример логлинейного анализа данных файла helpLLM.sav. В результате применения пошагового метода была выявлена наиболее оптимальная логлинейная модель, включающая два взаимодействия: помощьхусловие и помощьхсимпатияхагрессия. Рассмотрим анализ остатков на примере этой модели.

Для того чтобы в результате логлинейного анализа получить исчерпывающую информацию об остатках (см. главу 25), необходимо отказаться от пошагового метода и задать модель в диалоговом окне Loglinear Analysis: Model (Логлинейный анализ: Модель), а затем в диалоговом окне General Loglinear Analysis: Options (Общий логлинейный анализ: Параметры) установить флажок Residuals (Остатки) в группе Display (Отображать). После задания модели становятся доступными и флажки

в группе Plot (Диаграмма) — остается только установить нужные. Обратите внимание, что диаграммы не строятся для насыщенной модели (модели, включающей все главные факторы и взаимодействия) и пошагового метода.

В последней таблице раздела «Представление результатов» главы 25 имеется столбец Residual (Остаток). В этом столбце представлена информация об остатках, аналогичная той, которая была бы получена в случае принудительного задания модели. В этом столбце нет ячеек с очень высокими значениями остатков (максимум — 1,43).

Приведенная на рис. 26.3 диаграмма остатков, на первый взгляд, весьма запутана. На самом деле это изображение является аналогом корреляционной матрицы, в каждой ячейке которой находится диаграмма разброса. Как и в любой корреляционной матрице, содержимое ячеек, находящихся под главной диагональю, совпадает с содержимым ячеек, находящихся над главной диагональю. Диагональные ячейки представляют каждую строку и столбец. Так, средняя ячейка левого столбца является диаграммой разброса для ожидаемых и наблюдаемых частот ячеек, нижняя левая ячейка — диаграммой разброса наблюдаемых значений с остатками, а средняя ячейка в нижней строке — диаграммой разброса ожидаемых значений с остатками.



**Рис. 26.3.** Диаграмма остатков в виде корреляционной матрицы

К двум важным типам диаграмм относятся диаграмма ожидаемых и наблюдаемых частот, а также диаграмма ожидаемых частот и остатков. Для хорошей модели зависимость между наблюдаемыми и ожидаемыми значениями линейная или близкая к ней, а ожидаемые значения не зависят от остатков. Если это не так для исследуемой модели, то последнюю необходимо изменить, введя в нее дополнительную переменную либо исключив какое-либо из допущений, на которых основана модель.

В нашем примере зависимость между ожидаемыми и наблюдаемыми значениями очень близка к линейной. Кроме того, не наблюдается зависимости между ожидаемыми значениями и остатками, что свидетельствует о хорошей подгонке модели.

Нормальная (квантильная) диаграмма исправленных остатков, представленная на рис. 26.4, отображает фактические остатки по горизонтальной оси и ожидаемые нормально распределенные остатки по вертикальной оси. Хорошая модель должна показывать линейную зависимость двух указанных величин. Обратите внимание, что полученная кривая по виду напоминает кривую накопленных частот, приведенную ранее в этом разделе (см. рис. 26.1). Поскольку не наблюдается значительного отклонения этой зависимости от прямой линии, у исследователя не должно возникать беспокойств относительно неадекватности модели.

Normal Q-Q Plot of Standardized Residuals

Observed Value

**Рис. 26.4.** Нормальная (квантильная) диаграмма исправленных остатков

Диаграмма на рис. 26.5 является вариацией предыдущей. Разница заключается в том, что «идеальная» линия расположена не но диагонали, а проходит горизонтально; вертикальная ось представляет отклонение исследуемой кривой от идеальной линии.

Подобная форма представления позволяет увидеть отклонения от нормальных значений в увеличенном виде, что бывает удобно при исследовании кривых, близких к идеальным. Так, на показанной диаграмме четыре точки, находящиеся в правом нижнем углу, имеют большее отклонение, чем остальные, однако если принять во внимание масштаб вертикальной оси, эти отклонения весьма незначительны.

Графические средства анализа остатков демонстрируют для данного примера хорошую подгонку модели. Однако на практике такой результат встречается нечасто, и тогда требуется либо изменить модель, либо провести дополнительный анализ тех объектов, в отношении которых модель несостоятельна, то есть дает наибольшие величины остатков.

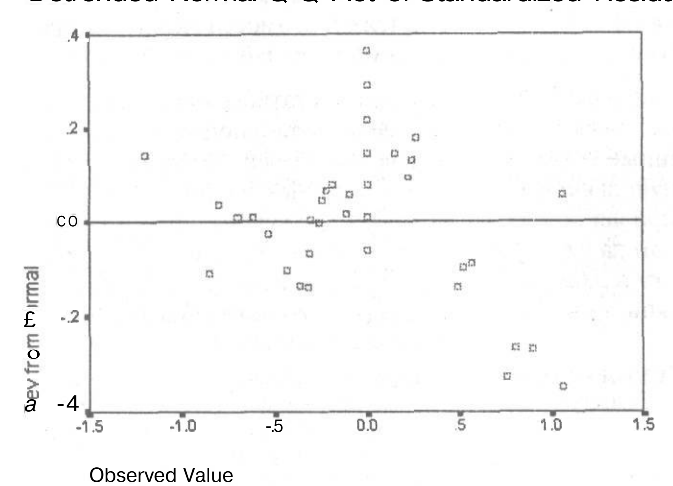



**Рис. 26.5.** Еще один вариант нормальной диаграммы исправленных остатков

# Инструменты SPSS для работы с остатками

Изложенный материал является лишь кратким введением в тему анализа остатков. Многие команды SPSS позволяют сохранять остатки в виде новых переменных файла данных, что делает возможным их всестороннее статистическое исследование.

В табл. 26.1 подводится итог темы анализа остатков, которой мы касались в различных главах этой книги. Для каждой главы указаны диалоговые окна и элементы интерфейса, связанные с анализом остатков.

**Таблица 26.1.** Главы книги, в которых упоминается анализ остатков

| Глава | Окно | Параметры | Описание |
|---|---|---|---|
| 8 | Crosstabs: Cell Display (Таблицы сопряженности: Ячейки), рис. 8.2 | Флажки Unstandardized (Нестандартизированные). Standardized (Стандартизованные) и Adj. Standardized (Исправленные стандартизованные) в группе Residuals (Остатки) | Вычисление остатков для каждой ячейки таблицы сопряженности. Стандартизованные остатки имеют среднее значение О, и стандартное отклонение 1; исправленные стандартизованные остатки определяют число стандартных отклонений, лежащих выше и ниже среднего значения |
| 15 | Multivariate: Options (Многомерный дисперсионный анализ: Параметры), рис. 15.5 | Флажки Residual Plots (Диаграммы остатков) и Residual SSCP Matrix (Матрица остатков) | Построение для каждой зависимой переменной диаграмм разброса, включающих ожидаемые, фактические и стандартизованные значения остатков, Остаточная сумма квадратов и матрица остатков позволяют выяснить; корреляции между остатками различных зависимых переменных |

| Глава | Окно | Параметры | Описание |
|---|---|---|---|
| 18 | Linear Regression: Save (Линейная регрессия: Сохранение), рис. 18.3 | Флажки Unstandardized (Нестандартизированные), Standardized (Стандартизованные), Studentized (Стьюдента). Deleted (Удаленные), Studentized deleted (Стьюдента удаленные) в группе Residuals (Остатки) | Сохранение различных типов остатков в файле данных в качестве новых переменных. Остатки Стьюдента стандартизованы на основе расстояния объекта от среднего значения. Удаленные остатки — это остатки, рассчитанные в предположении, что соответствующий объект исключен из анализа |
| 24 | Logistic Regression: Options (Логистическая регрессия: Параметры), рис. 24.3 | Флажок Casewise listing of residuals (Список остатков для объектов) | Перечисление остатков для всех объектов либо для тех объектов, которые лежат на расстоянии не менее заданного числа стандартных отклонений от среднего (по умолчанию 2) |
| | **Logistic** Regression: Save (Логистическая регрессия: Сохранение) | Все флажки в группе Residuals (Остатки) | Сохранение различных типов остатков в файле данных в качестве новых переменных. Остатки Стьюдента стандартизованы на основе расстояния объекта от среднего значения, остатки отклонения — на основе вероятности принадлежности объекта к группе, а логит-остатки — при помощи логит-шкал |
| 25 | Loglinear Analysis: Options (Логлинейный анализ: Параметры), рис. 25.3 | Флажок Residuals (Остатки): в группе Display (Отображать), а также оба флажка в группе Plot (Диаграмма) | Отображение остатков (в том числе стандартизованных), создание диаграмм остатков |

# Глоссарий

B. Показатели *B* представляют собой набор коэффициентов и константу регрессионного уравнения. Показатель *B* можно рассматривать как весовой коэффициент, характеризующий влияние соответствующей независимой переменной (предиктора) па зависимую переменную (критерий). Положительное значение *B* указывает на то, что с возрастанием предиктора значение критерия возрастает, а отрицательное значение *B* — па то, что значение критерия убывает.

**df.** *См. Число степеней свободы.*

***F*-критерий.** В дисперсионном анализе отношение межгруппового среднего квадрата к внутригрупповому среднему квадрату. Данная величина позволяет сравнить межгрупповую дисперсию с внутригрупповой дисперсией. В случае если первая окажется значительно выше второй, это будет означать наличие значимого различия между группами. В множественном регрессионном анализе F-критерий позволяет определить значимость множественной корреляции.

K. В иерархической логлинейной модели порядок взаимодействия эффектов; $k = 1$ соответствует первому порядку (одна переменная), $k = 2$ — второму порядку (две переменные), и т. д.

***p*-уровень.** *См. значимость.*

R. Множественный коэффициент корреляции между зависимой переменной и двумя или более независимыми переменными. Значение *R* лежит в пределах от 0 до 1 и интерпретируется но аналогии с обычным (двухмерным) коэффициентом корреляции.

**$R^2$.** Квадрат коэффициента множественной корреляции (коэффициент детерминации), доля дисперсии зависимой переменной, обусловленная воздействием двух или более независимых переменных.

**5-стресс.** В многомерном шкалировании мера степени соответствия модели исходной матрице различий. Чем меньше это значение, тем лучше соответствие.

***t*-критерий в регрессионном анализе.** Критерий, определяющий статистическую значимость корреляций, равен отношению коэффициента *B* к своей стандартной ошибке.

***t*-критерий для зависимых выборок.** Критерий, сравнивающий средние значения двух распределений для одной и той же выборки.

***t*-критерий для независимых выборок.** Критерий, сравнивающий средние значения одной и той же переменной для двух независимых выборок.

***t*-критерий для одной выборки.** Критерий, предназначенный для сравнения среднего значения распределения переменной с некоторой эталонной величиной.

***t*-критерий.** Критерий для определения статистической значимости различия двух средних.

**V Крамера.** Мера ассоциации между значениями двух категориальных переменных. Значение V всегда варьирует от 0 до 1 и интерпретируется по аналогии с коэффициентом корреляции (исключая отсутствие отрицательных значений). Нередко используется в контексте $\chi^2$-анализа; вычисление осуществляется по формуле (*k* — наименьшее из количеств строк и столбцов):

$$V = \sqrt{\chi / [N(k-1)]}.$$

**z-значения.** Также называются стандартизованными значениями. После стандартизации (или *z*-преобразования) значений переменной среднее равно 0, стандартное отклонение равно 1. Стандартизованное значение может характеризовать направление и степень отклонения исходного значения от среднего. Для стандартизованных значений, превышающих по модулю 1,96, уровень значимости оказывается ниже 0,05.

**Альфа** (а). Мера внутренней согласованности измерительной шкалы, вычисляемая по формуле $a = rk/[1 + (k - 1)r]$, где *k* — число переменных в анализе, *r* — среднее значение корреляции между пунктами шкалы. Значение а зависит от числа переменных, поэтому нет точной интерпретации его величины; тем не менее в большинстве случаев действует следующая оценка внутренней согласованности шкалы: а > 0,9 — отличная; а > 0,8 — хорошая; а > 0,7 — приемлемая; а > 0,6 — сомнительная; а > 0,5 — малопригодная; а < 0,5 — недопустимая.

**Альфа, если элемент удален.** В анализе надежности значение а для шкалы, получающейся путем удаления текущего ее пункта.

**Априорная вероятность для каждой группы.** Вероятность, характеризующая предполагаемое соотношение численности групп. Для каждой из двух групп она равна 0,5, если предполагается, что их численность одинакова.

**Асимметричная матрица.** Квадратная матрица, у которой хотя бы в одной паре ячеек, симметрично расположенных относительно главной диагонали, значения различны. Корреляционная матрица является типичным примером асимметричной матрицы.

**Асимметрия.** Мера отклонения распределения от нормального, характеризующая симметричность графика.

**Асимптотические значения.** Величины, на которых основано определение оценок параметров. Оценки параметров вычисляются в случаях, когда определение точных значений невозможно, в частности, в регрессионном анализе и некоторых других статистических процедурах.

**Барлетта критерий сферичности.** Критерий многомерной нормальности для распределения переменных. Помимо нормальности критерий проверяет, отличаются ли корреляции от 0. Значение *р*-уровня, меньшее 0,05, указывает на то, что данные вполне приемлемы для проведения факторного анализа

**Бета (β).** В регрессионном анализе Р означает стандартизованный коэффициент регрессии и представляет собой *В*-коэффициент для нормализованных переменных. Значения (3 всегда лежат в интервале от -1 до +1 и могут сравниваться друг с другом для разных переменных.

**Бета при включении.** Данная величина используется во множественном регрессионном анализе для переменных, не вошедших в уравнение регрессии, и представляет собой значение коэффициента р, рассчитанного в предположении, что соответствующая переменная была включена в регрессионное уравнение.

**Биномиальный критерий.** Непараметрический критерий, определяющий степень близости эмпирического распределения бинарной переменной к биномиальному распределению, для которого частоты двух категорий равны.

**Бонферрони критерий.** Критерий для множественного сравнения средних, если в дисперсионном анализе получен значимый результат.

**Вальда критерий.** В модели логистической регрессии критерий значимости коэффициента *В* для соответствующего предиктора. Чем выше его значение (вместе с числом степеней свободы), тем выше значимость

**Вероятность.** Ожидаемая относительная частота некоторого события.

**Взаимодействие.** Эффект совместного влияния на зависимую переменную двух и более независимых переменных, который не сводится к их раздельному влиянию. В случае двух независимых переменных проявляется в том, что эффект влияния одной из них проявляется по-разному на разных уровнях другой переменной.

**Вилкоксона критерий.** Непараметрический критерий, сходный с критерием знаков, однако в отличие от последнего использующий ранги положительных и отрицательных разностей.

**Внутригрупповая сумма квадратов.** Сумма квадратов отклонений наблюдаемых значений от среднего для каждой группы.

**Внутригрупповой многомерный дисперсионный анализ (MANOVA).** Вид дисперсионного анализа, в котором одна и та же группа объектов подвергается действию каждого уровня независимой неременной. То же самое, что и многомерный дисперсионный анализ с повторными измерениями.

**Вращение.** Процедура, применяемая в факторном анализе для того, чтобы получить более простую структуру факторов.

**Выборка.** Подмножество объектов из некоторой генеральной совокупности, выбранное для статистических выводов относительно свойств всей совокупности.

**Гистограмма.** Столбиковая диаграмма для отображения распределения частот по категориям (диапазонам значений) переменной. Горизонтальная ось графика соответствует значениям переменной, а вертикальная — частотам.

**Главный эффект.** Воздействие независимой переменной на зависимую переменную. Примеры главных эффектов можно найти в главах 14 и 15.

**График собственных значений.** Диаграмма, позволяющая выбрать число факторов в факторном анализе на основе критерия каменистой осыпи Р. Кеттелла.

**Гуттмана критерий половинного расщепления.** В анализе надежности половинного расщепления значение надежности, полученное с помощью процедуры нижних пределов.

**Дендрограмма.** Диаграмма древовидной структуры, иллюстрирующая процесс кластеризации в кластерном анализе. Пример дендрограммы приведен в главе 22.

**Детерминант ковариационно-дисперсионной матрицы.** Величина, характеризующая степень зависимости между значениями переменных. Чем меньше значение детерминанта, тем сильнее соответствующая зависимость. Эта величина используется при вычислении *M* Бокса. Детерминант общей дисперсионно-ковариационной матрицы учитывает все матрицы, используемые в анализе.

**Диаграмма последовательности слияния.** Графическая интерпретация пошаговой процедуры кластеризации в кластерном анализе.

**Диаграмма рассеивания.** График для анализа связи между двумя переменными, на котором каждый объект представляет собой точку. Положение точки задано парой значений двух переменных для данного объекта. Более подробное описание приведено в главе 9.

**Диаграмма регрессии.** Диаграмма разброса, включающая сдвиги точек от линии регрессии по вертикальной оси.

**Дискриминантный анализ.** Процедура создания формулы регрессии, на основе которой производится разбиение объектов на группы, соответствующие категориям зависимой переменной.

**Дисперсии элементов.** Аналог средних значений элементов (пунктов шкалы) в анализе надежности. Поясняющий пример приведен в разделе «Представление результатов» главы 19.

**Дисперсионный анализ (ANOVA).** Статистический анализ, устанавливающий статистическую значимость различий между средними значениями для трех или более выборок.

**Дисперсия.** Характеристика выборочного распределения переменной, описывающая разброс значений вокруг среднего и вычисляемая как отношение суммы квадратов отклонений к объему выборки, уменьшенному на 1. Кроме того, дисперсия представляет собой квадрат стандартного отклонения.

**Дисперсия шкалы, если элемент удален.** Дисперсия суммы всех пунктов шкалы, кроме удаленного пункта.

**Доверительный интервал.** Диапазон, в котором находится большинство значений выборки. Например, термин «доверительный интервал в 95 %» означает интервал, в который любое случайное значение из выборки попадает с вероятностью 95 %.

**Доверительный интервал в 95 %.** См. *Доверительный интервал*.

**Знаков критерий.** Непараметрический критерий, определяющий различие двух измерений для одной выборки на основе знаков разностей пар значений.

**Значимость ($p$-уровень).** Мера случайности полученного результата, равная вероятности того, что в генеральной совокупности этот результат (различия, связь) отсутствует. Чем меньше эта вероятность (значение $p$-уровня), тем выше статистическая значимость результата. Результат считается статистически достоверным (значимым), если р-уровень не превышает 0,05.

**Изменение $R^2$.** Изменение величины $R^2$ в результате введения новой переменной в уравнение регрессии.

**Исправленная величина $R^2$.** Во множественном регрессионном анализе величина $R^2$ является точной для выборок, однако в генеральной совокупности ее значение лишь приблизительно. Исправленная величина $R^2$ представляет собой более точную оценку $R^2$ для генеральной совокупности и используется при сравнениях моделей, содержащих различное число независимых переменных.

**Итерация.** Стадия процесса формирования регрессионного (дискриминантного) уравнения, на которой происходит включение или исключение очередной переменной. Процесс продолжается до тех пор, пока не перестанет удовлетворяться заданный в процедуре критерий.

**Кайзера—Мейера—Олкина критерий адекватности выборки.** Величина, характеризующая степень применимости факторного анализа к данной выборке. Правило интерпретации этого критерия следующее:

- более 0,9 — безусловная адекватность;
- более 0,8 — высокая адекватность;
- более 0,7 — приемлемая адекватность;
- более 0,6 — удовлетворительная адекватность;
- более 0,5 — низкая адекватность;
- менее 0,5 — факторный анализ неприменим к выборке.

**Канонические дискриминантные функции.** Одно или более линейное дискриминантное уравнение, построенное таким образом, что классификация объектов по уровням зависимой переменной происходит наиболее точно. Более подробное описание вы можете найти в главе 23.

**Канонические коэффициенты.** В дискриминантном анализе канонический коэффициент представляет собой корреляцию между оценками дискриминантной функции и уровнями зависимой переменной. Более подробное описание вы можете найти в главе 23.

**Категориальная (номинативная) переменная.** Переменная, каждое значение которой указывает на принадлежность объекта к определенной группе (категории). Категориальная переменная не является количественной; она разделяет все объекты на непересекающиеся группы по определенному признаку (пол, хобби, класс и пр.), но не позволяет сравнивать объекты по уровню выраженности этого признака.

**Квадрат евклидового расстояния.** Мера, используемая по умолчанию в кластерном анализе для определения расстояния между объектами и кластерами и вычисляемая как сумма квадратов разностей между значениями переменных двух объектов.

**Квадрат эта ($\eta^2$).** Доля дисперсии зависимой переменной, обусловленная воздействием со стороны независимой переменной. Так, $\eta^2 = 0{,}044$ означает, что 4,4 % дисперсии зависимой переменной обусловлено данной независимой переменной.

**Квадратная матрица.** Матрица, строки и столбцы которой соответствуют одной и той же последовательности элементов (переменных или объектов).

**Квартили.** 25, 50 и 75-й процентили.

**Кластерный анализ.** Процедура, на основе заданного правила объединяющая объекты или переменные в группы, называемые кластерами.

**Ковариата.** Количественная переменная, имеющая значительную корреляцию с зависимой переменной и включаемая в анализ для более точной проверки воздействий факторов на зависимую переменную.

**Количественная переменная.** Значения количественной переменной (в отличие от категориальной) отражают уровень выраженности у объектов соответствующего признака в метрической или порядковой шкале.

**Колмогорова—Смирнова критерий для одной выборки.** Непараметрический критерий, определяющий, отличается ли данное эмпирическое распределение от теоретического распределения (нормального, равномерного, Пуассона или экспоненциального).

**Контрасты.** Метод контрастов — это метод множественного сравнения средних в дисперсионном анализе, который позволяет сравнивать выборки по градациям независимой переменной. Например, контрасты позволяют сравнивать одну градацию с другой, одну градацию со всеми остальными или разбить все градации на 2 группы и затем сравнить их между собой.

**Корреляция.** Мера степени и направления связи между значениями двух переменных. Понятию корреляции посвящена глава 9.

**Корреляция между формами.** В анализе надежности половинного расщепления приближенное значение надежности измерения в предположении, что обе половины содержат одинаковое число пунктов.

**Корреляция между элементами.** В анализе надежности это описательная информация о корреляциях каждого пункта с суммой всех остальных пунктов.

**Коэффициент корреляции.** Мера связи двух переменных, обозначаемая символом $r$ и принимающая значения от -1 до +1.

**Коэффициенты регрессии.** $B$-коэффициенты, то есть множители при переменных, входящих в состав регрессионного уравнения, а также константа.

**Критерий согласия.** В логлинейном анализе критерий $\chi^2$ для определения степени адекватности модели исходным данным. Чем выше его значения и чем ниже соответствующие уровни значимости, тем хуже модель соответствует данным.

**Левина критерий.** Критерий, предназначенный для проверки гипотезы о том, что все распределения зависимой переменной для сравниваемых выборок имеют одинаковые дисперсии.

**Линия регрессии.** Прямая линия на графике двухмерного рассеивания, отражающая наиболее точные прогнозируемые значения («линия наилучшего соответствия»).

**Логарифмический детерминант.** В дискриминантном анализе натуральный логарифм определителя каждой ковариационной матрицы. Логарифмический определитель используется для вычисления *M* Бокса.

**Логит.** Натуральный логарифм шанса. Описание понятия «логит» приводится в главе 24.

**Лямбда Уилкса.** Отношение внутригрупповой суммы квадратов к общей сумме квадратов, характеризующее дисперсию оценок дискриминантной функции, не обусловленную различиями между двумя группами. Единичное значение лямбда принимает в случае, если наблюдаемые средние значения групп равны; значения, близкие к нулю, означают, что внутригрупповая дисперсия мала по сравнению с общей дисперсией.

***M* Бокса *(M)*.** Критерий многомерной нормальности, основанный на близости значений определителей матриц ковариаций двух или более групп.

**Максимум.** Наибольшее наблюдаемое значение распределения переменной.

**Манна–Уитни и Вилкоксона критерий ранговых сумм.** Непараметрический аналог $t$-критерия, определяющий различие между двумя выборками на основе рангов.

**Матрица различий.** Матрица, каждое значение которой соответствует различию между двумя объектами.

**Матрица трансформации факторов.** Результатом умножения этой матрицы и матрицы факторных нагрузок до вращения является матрица нагрузок факторов после вращения.

**Медиана.** Значение переменной, делящее упорядоченное множество всех значений выборки ровно пополам: у половины объектов выборки значения переменной больше, а у другой половины меньше медианы.

**Медианный критерий *k* выборок.** Непараметрический критерий для сравнения двух и более выборок по уровню выраженности признака; основан на подсчете числа элементов, лежащих выше и ниже главной медианы.

**Межгрупповая сумма квадратов.** Сумма квадратов разностей между главным средним значением и средними значениями групп, умноженных на весовые коэффициенты, равные числу объектов в соответствующих группах.

**Метод главных компонентов.** Метод, применяемый SPSS по умолчанию в факторном анализе для извлечения факторов.

**Метод иерархического слияния.** Метод, используемый в кластерном анализе, при выполнении которого объекты объединяются в кластеры по одному на каждом шаге до тех пор, пока не будет образован единственный кластер, охватывающий все объекты. Кластерный анализ подробно описан в главе 22.

**Метрическая переменная.** Количественная переменная, соответствующая измерению признака в шкале интервалов или отношений. В отличие от ранговой (порядковой) переменной, при сравнении объектов позволяет судить не только о том, больше или меньше выражен признак, но и о том, насколько больше (меньше) он выражен.

**Минимум.** Наименьшее наблюдаемое значение распределения переменной.

**Многомерное шкалирование.** Метод, позволяющий на основе матрицы различий между объектами построить одно-, двух- или трехмерное изображение, иллюстрирующее удаленность этих объектов друг от друга.

**Многомерные критерии значимости.** В многомерном дисперсионном анализе набор критериев, позволяющих определить влияние факторов и их взаимодействий на совокупность зависимых переменных. Наиболее мощным считается критерий Пилая.

**Многомерный дисперсионный анализ (MANOVA).** Отличие многомерного дисперсионного анализа от одномерного (ANOVA) заключается в том, что число зависимых переменных в нем может быть теоретически любым.

**Многомерный дисперсионный анализ с повторными измерениями.** Вид дисперсионного анализа, в котором одна и та же группа объектов подвергается действию каждого уровня независимой переменной. С точки зрения вычислений этот анализ можно назвать внутригрупповым.

**Многомерный ковариационный анализ (MANCOVA).** Многомерный дисперсионный анализ с включением в анализ ковариат.

**Многомерный критерий однородности матриц ковариаций.** Критерий *M* Бокса определяет, являются ли ковариационные матрицы одинаковыми. Для каждого из значений вычисляется *p*-уровень, а также величина *F* или $\chi^2$.

**Множественный регрессионный анализ.** Метод, позволяющий спрогнозировать значения зависимой переменной на основе известных значений независимых переменных.

**Мода.** Наиболее часто повторяющееся значение распределения переменной.

**Моучли критерий сферичности.** Критерий многомерной нормальности. SPSS вычисляет приблизительное значение $\chi^2$ и соответствующий уровень значимости. Если уровень значимости оказывается менее 0,05, то, вероятно, данные не являются нормально распределенными.

**Наблюдаемое значение или частота.** В $\chi^2$-анализе фактическая частота категории.

**Надежность половинного расщепления.** Мера надежности, для вычисления которой все пункты шкалы делятся на две эквивалентные группы, а затем на основе корреляции между двумя половинами шкалы устанавливается ее внутренняя согласованность.

**Наименьшая ожидаемая частота.** Наименьшая частота в ячейке таблицы сопряженности, определяемая в процессе применения критерия $x^2$.

**Наименьшей значимой разности критерий.** Критерий множественного сравнения средних, представляющий собой серию $t$-критериев; применяется, если в дисперсионном анализе получен значимый результат.

**Накопленная частота.** Суммарное число объектов, имеющих значение переменной, не большее, чем указано.

**Накопленный процент.** Процент объектов от общего числа, имеющих значение переменной, не большее, чем указано.

**Насыщенная модель.** Логлинейная модель, включающая все взаимодействия и главные эффекты факторов.

**Нелинейная регрессия.** Процедура вычисления параметров нелинейного регрессионного уравнения.

**Неортогональное вращение.** Процедура, используемая в факторном анализе, допускающая результат, в котором угол между факторами отклоняется от прямого. Это иногда желательно для достижения более простой структуры.

**Непараметрические критерии.** Серия критериев, каждый из которых применяется без предварительных допущений относительно нормальности распределения. Непараметрические критерии основаны на ранжировании, попарных сравнениях и других средствах, не требующих нормальности распределения переменных.

**Нестандартизированные коэффициенты канонической дискриминантной функции.** Список коэффициентов и константа дискриминантного уравнения.

**Номинативная шкала.** *См. Категориальная переменная.*

**Нормальное распределение.** Распределение частот (вероятностей), графически представляемое в виде симметричной кривой, имеющей пик в центре и асимпто-

тически приближающееся к горизонтальной оси по краям. Идеальное нормальное распределение характеризуется нулевыми значениями асимметрии и эксцесса.

**Общая внутригрупповая ковариационная матрица.** Матрица, состоящая из средних значений ковариационных матриц, вычисленных для каждого уровня зависимой переменной.

**Общая сумма квадратов.** Сумма квадратов отклонений всех значений от среднего значения всего распределения.

**Общность.** В факторном анализе мера, характеризующая долю дисперсии переменной, обусловленную воздействием всех факторов.

**Одномерные *F*-критерии.** В многомерном дисперсионном анализе критерии, которые характеризуют влияние независимых переменных и их взаимодействий на каждую зависимую переменную в отдельности.

**Ожидаемое значение.** В перекрестной таблице при использовании критерия $x^2$ значение, вычисляемое в предположении, что все переменные являются полностью независимыми друг от друга. В регрессионном анализе термин «ожидаемое значение» эквивалентен термину «прогнозируемое значение» и означает величину, получаемую для каждого объекта в результате подстановки значений переменных для него в уравнение регрессии.

**Остатки и стандартизованные остатки.** В логлинейных моделях остатки представляют собой разности между ожидаемыми и наблюдаемыми частотами. Чем выше значения остатков, тем менее адекватной является модель. SPSS подсчитывает исправленные величины остатков с использованием оценок стандартного отклонения. Исправленные остатки представлены в единицах нормального распределения, и, как правило, значения остатков, по модулю превышающие 1,96, являются значимыми.

**Остаток.** Как правило, разность между наблюдаемым и ожидаемым значениями. Эта величина относится к части дисперсии, которая не объясняется воздействием независимых переменных.

**Отклонение.** Расстояние и направление (отрицательное или положительное) между средним и данным значениями.

**Оценка дискриминантной функции.** Значение, получаемое для каждого объекта путем подстановки значений его переменных в уравнение дискриминантной функции.

**Параметр.** Некоторая числовая характеристика генеральной совокупности.

**Параметрические критерии.** Критерии, применяемые в предположении о нормальном распределении переменных в генеральной совокупности.

**Переменные в уравнении.** При выводе результатов пошагового регрессионного анализа SPSS включает для каждого шага статистики тех переменных, которые вошли в уравнение регрессии.

**Пирсона коэффициент корреляции.** Мера корреляции, идеально подходящая для двух непрерывных (метрических) переменных.

**Порядковая (ранговая) шкала.** *См. Ранговая переменная.*

**Пошаговый выбор переменных.** Процедура, включающая и исключающая переменные из дискриминантного или регрессионного уравнения в соответствии с выбранными критериями.

**Прямоугольная матрица.** Матрица, для которой строкам и столбцам соответствуют разные последовательности элементов (объектов или переменных).

**Размах.** Характеристика распределения, равная разности между минимумом и максимумом распределения.

**Ранговая (порядковая) переменная.** Количественная переменная, отражающая измеренное качество на уровне порядка: в большей или меньшей степени оно выражено. В отличие от метрической шкалы не позволяет судить о том, насколько больше или меньше выражено качество, поэтому не допускает применения арифметических операций.

**Распределение.** Статистическое понятие, обозначающее соотношение значений признака и частот (вероятностей) их встречаемости. Распределение (вероятностей, частот) может быть представлено в виде формулы для функции распределения вероятностей, графика распределения частот (гистограммы, столбиковой диаграммы), таблицы распределения частот.

**Регрессионный анализ.** Инструмент статистики, позволяющий прогнозировать значения зависимой переменной с помощью известных значений независимых переменных. Подробное рассмотрение регрессионного анализа проводится в главах 17 и 18.

**Регрессия.** В множественном регрессионном анализе этим термином обозначается статистика, отражающая влияние предикторов на зависимую переменную.

**Серий критерий.** Непараметрический критерий, определяющий, является ли последовательность бинарных величин (событий) случайной или упорядоченной.

**Симметричная матрица.** Квадратная матрица, для которой в каждой паре ячеек, расположенных симметрично относительно главной диагонали, содержатся одинаковые значения. Типичным примером симметричной матрицы является корреляционная матрица.

**Скорректированная корреляция пункта и суммы.** В анализе надежности корреляция между пунктом шкалы и суммой всех остальных пунктов.

**Собственное значение.** В факторном анализе эта величина пропорциональна доле дисперсии, обусловленной влиянием данного фактора; в дискриминантном анализе отношение межгрупповой суммы квадратов к внутригрупповой сумме квадратов. Чем больше собственное значение, тем выше точность дискриминантной функции.

**Спирмена—Брауна критерий неэквивалентных форм.** В анализе надежности половинного расщепления — надежность, вычисленная для случая, когда «половины» имеют неравный размер.

**Спирмена—Брауна критерий эквивалентных форм.** Используется в анализе надежности, когда число элементов в «половинах» одинаково (вычисляется коэффициент корреляции).

**Среднее шкалы, если элемент удален.** В анализе надежности для каждого пункта шкалы вычисляется сумма остальных пунктов по всем объектам выборки; отношение указанной суммы к числу объектов является средним шкалы, если данный элемент удален.

**Средние значения элементов.** В анализе надежности (с применением альфа Кронбаха) — описательная информация, касающаяся средних значений пунктов шкалы по всем объектам. Поясняющий пример приведен в разделе «Представление результатов» главы 19.

**Средний квадрат.** Отношение суммы квадратов к числу степеней свободы. В однофакторном дисперсионном анализе, как правило, средний квадрат вычисляется для внутригрупповой и межгрупповой сумм квадратов, а в регрессионном анализе — для регрессионной и остаточной сумм квадратов. Во всех перечисленных случаях средний квадрат используется для вычисления $F$-критерия.

**Стандартизованный коэффициент** а. В анализе надежности — значение $\alpha$, полученное в случае, если перед проведением анализа стандартизовать распределения всех элементов шкалы.

**Стандартная ошибка.** Стандартное отклонение величины, получаемое в результате ее многократного вычисления для случайных выборок. Как правило, стандартная ошибка вычисляется для среднего значения распределения.

**Стандартное отклонение.** Мера разброса значений распределения вокруг среднего. Стандартное отклонение определяется как квадратный корень дисперсии (суммы квадратов отклонений от среднего, деленной *на* $N - 1$, где $N$ — объем выборки).

**Статистики для сумм переменных.** В анализе надежности статистические характеристики суммы всех переменных по объектам.

**Столбиковая диаграмма.** График распределения частот по категориям (значениям) переменной. Каждый столбец на графике соответствует одному значению признака, а его высота пропорциональна частоте встречаемости этого значения. Аналогичное средство для количественных переменных, имеющих большое число возможных значений, обычно называется гистограммой.

**Стресс.** В многомерном шкалировании мера соответствия модели исходной матрице различий. Чем меньше значение стресса, тем лучше соответствие модели.

**Сумма квадратов.** Стандартная мера разброса, представляющая собой сумму квадратов отклонений всех значений величины от среднего значения.

**Таблица распределения (частот).** Таблица, устанавливающая соотношение между категориями (значениями) признака и частотами их встречаемости.

**Таблица сопряженности (кросстабуляции).** Обычно таблица совместного распределения частот для двух категориальных или дискретных переменных; строки соответствуют категориям (значениям) одной, а столбцы — другой переменной. Подробно таблицы сопряженности описаны в главе 8.

**Толерантность.** Мера линейной зависимости между одной переменной и набором других переменных. Если при дискриминантном анализе уровень толерантности составляет менее 0,001, это означает, что линейная зависимость для данной переменной настолько высока, что ее включение в дискриминантное уравнение недопустимо.

**Тьюки критерий подлинной значимости.** Критерий множественного сравнения, позволяющий попарно сравнивать средние значения; применяется, если дисперсионный анализ показал значимый результат.

**Фактор.** В факторном анализе объединение нескольких переменных, чья взаимная корреляция исчерпывает определенную долю общей дисперсии. После процедуры вращения каждый фактор интерпретируется как некоторая общая причина взаимосвязи группы переменных.

**Факторный анализ.** Метод, позволяющий свести большое количество исходных переменных к значительно меньшему числу факторов, каждый из которых объединяет исходные переменные, имеющие сходный смысл.

**Фи** ($\varphi$). Мера связи (корреляции) двух категориальных переменных, обычно применяемая наряду с критерием $x^2$ при анализе таблиц сопряженности и вычисляемая по формуле:

$$\varphi = \sqrt{\chi^2 / N}.$$

**Фридмана дисперсионный анализ.** Непараметрическая процедура, определяющая, различаются ли между собой три или более измерения для одной и той же выборки, на основе среднего ранга каждого измерения.

**Хи-квадрат ($\chi^2$) для модели.** В анализе логистической регрессии величина, позволяющая определить, оказывают ли переменные, входящие в состав регрессионного уравнения, значимое влияние на зависимую переменную. Чем выше полученное значение, тем значительнее воздействие.

**Хи-квадрат ($\chi^2$) критерий для одной выборки.** Непараметрический критерий, определяющий отличие наблюдаемого распределения переменной от ожидаемого (теоретического) распределения.

**Хи-квадрат критерий (критерий $\chi^2$).** Непараметрический критерий для сравнения ожидаемых и наблюдаемых частот (как правило, для таблиц сопряженности). Критерий $\chi^2$ может использоваться для оценки адекватности структурных

и логлинейных моделей. В любом случае, $\chi^2$-анализ всегда отвечает на один и тот же вопрос: отличаются ли ожидаемые частоты модели от наблюдаемых. Коэффициент $\chi^2$ Пирсона вычисляется по следующей формуле:

$$\chi^2 = \sum[(f_0 - f_e)^2 / f_e].$$

**Центроиды групп.** В дискриминантном анализе средние значения дискриминантных функций для каждой из двух или более групп. Если число групп равно 2, центроиды будут иметь одинаковые абсолютные величины и разные знаки. Чем ближе объект в центроиду группы, тем больше вероятность, что он принадлежит к этой группе.

**Частичный критерий $\chi^2$.** Значение критерия $\chi^2$, характеризующее долю воздействия очередной независимой переменной на зависимую переменную.

**Частота (абсолютная).** Количество объектов в выборке, имеющих данное значение признака.

**Частота относительная.** Доля объектов в выборке, имеющих данное значение признака; равна отношению абсолютной частоты к объему выборки.

**Число степеней свободы (df).** Количество возможных направлений изменчивости статистического показателя, наряду с эмпирическим значением критерия служит для определения p-уровня значимости.

**Шанс.** Отношение вероятности того, что событие произойдет, к вероятности того, что событие не произойдет.

**Шеффе критерий.** Процедура, позволяющая осуществлять попарные множественные сравнения средних значений после получения статистически достоверного результата дисперсионного анализа.

**Экспонента *B*.** В логистическом регрессионном анализе величина e*B* используется в одной из форм регрессионного уравнения и позволяет создать одну из интерпретаций коэффициентов регрессии.

**Эксцесс.** Мера «сглаженности» («островершинности» или «плосковершинности») распределения. Если значение эксцесса близко к 0, это означает, что форма распределения близка к нормальному виду. Детальное описание эксцесса приведено в главе 7.

**Эта** *(η).* Мера корреляции между двумя переменными в случае, если одна из них является категориальной.

# Англо-русский словарь терминов

В представленной таблице перечислены основные термины, с которыми может встретиться пользователь программы SPSS (в частности в окне вывода). Более подробную информацию об этих терминах можно найти в разделе «Представление результатов» каждой главы.

| Термин | Перевод |
|---|---|
| Advanced models unit | Модуль дополнительных моделей |
| Agglomeration schedule | Последовательность слияния |
| Alignment | Выравнивание |
| Alpha | Альфа |
| Analysis Of Variances (ANOVA) | Дисперсионный анализ |
| Approximate significance | Приблизительная значимость |
| Asymptotic significance | Асимптотическая значимость |
| Bar graph | Столбиковая диаграмма |
| Barlett's test of sphericity | Критерий сферичности Барлетта |
| Base system unit | Основной системный модуль |
| Between-groups linkage | Межгрупповое связывание |
| Binomial test | Биномиальный критерий |
| Box plot | Коробчатая диаграмма |
| Box's M | M Бокса |
| Canonical analysis | Канонический анализ |
| Category (x) axes | Ось категорий (x) |
| Category axes label | Метка оси категорий |
| Category axes title | Заголовок оси категорий |
| Cell | Ячейка (таблицы) |
| Central tendency | Центральная тенденция (мера) |
| Centroid | Центроид |
| Chi-square | Хи-квадрат |
| Chronbach's alpha | Альфа Кронбаха |
| Cluster analysis | Кластерный анализ |
| Cluster method | Метод кластеризации |
| Coefficient | Коэффициент |
| Commonality | Общность |
| Comparison | Сравнение |
| Complete linkage | Полная связь |
| Advanced models unit | Модуль дополнительных моделей |

| Термин | Перевод |
|---|---|
| Agglomeration schedule | Последовательность слияния |
| Alignment | Выравнивание |
| Alpha | Альфа |
| Analysis Of Variances (ANOVA) | Дисперсионный анализ |
| Approximate significance | Приблизительная значимость |
| Asymptotic significance | Асимптотическая значимость |
| Bar graph | Столбиковая диаграмма |
| Barlett's test of sphericity | Критерий сферичности Барлетта |
| Base system unit | Основной системный модуль |
| Between-groups linkage | Межгрупповое связывание |
| Binomial test | Биномиальный критерий |
| Box plot | Коробчатая диаграмма |
| **Box's M** | M Бокса |
| Canonical analysis | Канонический анализ; |
| Category (x) axes | Ось категорий (*x*) |
| Category axes label | Метка оси категорий |
| Category axes title | Заголовок оси категорий |
| Cell | Ячейка (таблицы) |
| Central tendency | Центральная тенденция (мера) |
| Centroid | Центроид |
| Chi-square | Хи-квадрат |
| Chronbach's alpha | Альфа Кронбаха |
| Cluster analysis | Кластерный анализ |
| Cluster method | Метод кластеризации |
| Coefficient | Коэффициент |
| Commonality | Общность |
| Comparison | Сравнение |
| Complete linkage | Полная связь |
| Confidence interval | Доверительный интервал |
| Continuity correction | Поправка на непрерывность |
| **Contrast** | Контраст |
| Convergence | Сходимость |
| **Correlation** | Корреляция |
| Correlation matrix | Корреляционная матрица |
| Count measure | Количественный показатель (мера) |
| Covariance | Ковариация |
| Covariate | **Ковариата** |
| Cox & Snell's R Square | *R*-квадрат Кокса и Снелла |
| Cramer's V | V Крамера |
| Criteria (criterion) | Условие |
| Crosstabulation (crosstab) | Таблица сопряженности (кросстабуляции) |

| Термин | Перевод |
|---|---|
| Cumulative frequencies | Кумулятивные (накопленные) частоты |
| Curve | Кривая (линия) |
| Curvilinear analysis | Криволинейный анализ |
| Cut point | Точка деления |
| Data reduction | Сокращение данных |
| Degrees of Freedom (df) | Число степеней свободы |
| Deletion | Удаление (исключение) |
| Dendrogram | Дендрограмма |
| Density function | Функция плотности вероятности |
| Dependent sample | Зависимая выборка |
| Descriptive statistic | Описательная статистика |
| Determinant | Детерминант (определитель) матрицы |
| Deviation | Отклонение |
| Difference | Разность, различие |
| Dimension | Шкала |
| Directory | Каталог |
| Discriminant analysis | Дискриминантный анализ |
| Dispersion | Изменчивость |
| Dissimilarity | Различие |
| Distance | Расстояние |
| Distribution | Распределение |
| Distribution function | Функция распределения: вероятности |
| Eigenvalue | Собственное значение |
| Enter | Включение |
| Epsilon corrected | С эпсилон-коррекцией |
| Equality of variances | Равенство дисперсий |
| Equal-length Spearman—Brown | Коэффициент эквивалентных форм Спирмена—Брауна |
| Equation | Уравнение |
| Error | Ошибка |
| Error bar chart | Диаграмма столбцов ошибок |
| Eta Squared | Квадрат эта |
| Euclidian distance | Евклидово расстояние |
| Expected frequency | Ожидаемая (теоретическая) частота |
| Extraction method | Метод экстракции |
| Factor | Фактор |
| Factor analysis | Факторный анализ |
| Factor score | Факторная оценка |
| Factor score coefficient | Коэффициент факторной оценки |
| Fisher's exact test | Точный критерий Фишера |
| Fixed factor | Постоянный фактор |
| Footnote | Сноска |

| Термин | Перевод |
|---|---|
| **Frequency** | Частота |
| Frequency table | Таблица частот |
| Friedman test | Критерий Фридмана |
| F-value | F-критерий |
| General linear model | Общая линейная модель |
| Graph | График |
| Grid line | Линия сетки |
| Group centroid | Центроид групп |
| Group membership | Принадлежность к группе |
| Grouping variable | Группирующая переменная |
| Guttman Split-Half | Критерий надежности половинного расщепления Гуттмана |
| Hierarchical cluster analysis | Иерархический кластерный анализ |
| Histogram | Гистограмма |
| Homogeneity of variances | Гомогенность (однородность) дисперсий |
| Honestly Significant Difference (HSD) | Критерий подлинной значимости |
| Hotelling T-square | Т-квадрат Хотеллинга |
| Icicle | Диаграмма накопления (слияния) |
| Image factoring | Факторный анализ образов |
| Improvement | Улучшение |
| Independent sample | Независимая выборка |
| Individual differences model | Модель индивидуальных различий |
| Initial Eigenvalue | Исходное собственное значение |
| Inner frame | Внутренняя рамка |
| Interaction | Взаимодействие |
| Inverse distribution function | Обратная функция распределения |
| Item | Пункт (элемент шкалы) |
| Item variance | Дисперсия элементов |
| Iteration | Итерация |
| Kaiser—Meyer—Olkin Measure of Sampling Adequacy | Критерий адекватности выборки Кайзера—Мейера—Олкина |
| Kendall's tau | Тау Кендалла |
| Kendall's tau-b | Тау-b Кендалла |
| KMO and Barlett test of sphericity | Критерии КМО и сферичности Барлетта |
| Kolmogorov—Smirnov test | Критерий Колмогорова—Смирнова |
| Kruskal—Wallis one-way analysis of variance | Однофакторный дисперсионный анализ Краскала—Уоллеса |
| Kurtosis | Эксцесс |
| Legend | Легенда (условные обозначения) |
| Legend label | Метка легенды |
| Level of measurement | Уровень (шкала) измерения |
| Levene's test | Критерий Левина |

| Термин | Перевод |
|---|---|
| Likelihood | Правдоподобие |
| Line (style: dotted) | Пунктирная линия |
| Line (style: solid) | Сплошная линия |
| Line graph | Линейный график |
| Linear regression analysis | Линейный регрессионный анализ |
| Linear-by-linear association | Линейная связь между переменными |
| Linkage | Соединение, связь |
| Listwise | Построчно |
| Loading | Нагрузка |
| Logistic regression | Логистическая регрессия |
| Logit | Логит |
| Loglinear analysis | Логлинейный анализ |
| Mahalanobis distance | Расстояние Махаланобиса |
| Main effect | Главный эффект |
| Mann—Whitney U | Критерий U Манна—Уитни |
| Marker | Маркер |
| Matrix | Матрица |
| Mauchly's test of sphericity | Критерий сферичности Моучли |
| Maximum | Максимум |
| McNemar test | Критерий Мак-Нимара |
| Mean | Среднее значение |
| Mean difference | Разность средних |
| Mean of Squares (MS) | Средний квадрат |
| Mean Rank | Средний ранг |
| Mean Square | Средний квадрат |
| Median | Медиана |
| Minimum | Минимум |
| Minimum expected count | Минимальная ожидаемая частота |
| Mode | Мода |
| Model selection | Подбор модели |
| Multidimensional scaling | Многомерное шкалирование |
| Multiple comparisons | Множественные сравнения (средних) |
| Multivariate Analysis Of Covariance (MANCOVA) | Многомерный ковариационный анализ |
| Multivariate Analysis Of Variances (MANOVA) | Многомерный дисперсионный анализ |
| Nagelkerke R Square | *R*-квадрат Нейджелкерка |
| No correlation | Отсутствие корреляции |
| Nonparametric test | Непараметрический критерий |
| Normal distribution | Нормальное распределение |
| Oblique rotation | Облическое (не ортогональное) вращение |
| Observed frequency | Наблюдаемая (эмпирическая) частота |
| One sample | Одна выборка |

| Термин | Перевод |
|---|---|
| One-sample Kolmogorov—Smirnov test | Критерий Колмогорова—Смирнова для одной выборки |
| One-tailed significance | Односторонняя значимость |
| Outer frame | Внешняя рамка |
| Paired samples | Зависимые выборки |
| Pairwise | Попарно |
| Parametric test | Параметрический критерий |
| Pareto chart | Парето-диаграмма |
| Partial correlation | Частная корреляция |
| Partial Eta Squared | Частичный квадрат эта |
| Pattern | Образец |
| Pearson Chi-Square | Критерий хи-квадрат Пирсона |
| Pearson correlation | Корреляция Пирсона |
| Percentile | Процентиль |
| Perfect negative correla[redacted] | Строгая отрицательная корреляция |
| Perfect positive correlation | Строгая положительная корреляция |
| Phi | Фи |
| Pie chart | Круговая диаграмма |
| Pillai's trace | След Пиллая |
| Plot | Диаграмма |
| Principal axis factoring | Факторный анализ методом главных осей |
| Principle component analysis | Анализ главных компонентов |
| Prior probability | Априорная вероятность: |
| Probability | Вероятность |
| Probability of group membership | Вероятность принадлежности к группе |
| Proximity | Близость |
| Quadratic relationship | Квадратичная связь (зависимость) |
| Quantile | Квантиль |
| Quartile | Квартиль |
| R Square | Коэффициент детерминации (квадрат коэффициента корреляции) |
| R Square change; | Изменение коэффициента детерминации |
| *Random* factor | Случайный фактор |
| Range | Размах |
| Rank | Ранг |
| Rectangular matrix | Прямоугольная матрица |
| Reference line | Вспомогательная линия |
| Regression coefficient | Коэффициент регрессии |
| **Regression** line | Линия регрессии |
| Regression model unit | Модуль регрессионных моделей |
| Related sample | Зависимая (связанная) выборка |
| Reliability analysis | Анализ надежности |

| Термин | Перевод |
|---|---|
| Remove method | Метод исключения |
| Repeated-measures ANOVA | Анализ ANOVA с повторными измерениями |
| Residual | Остаток |
| Residual analysis | Анализ остатков |
| Rotation method | Метод вращения |
| Sample | Выборка |
| Saturated model | Насыщенная модель |
| Scale (y) axes | Ось значений (y) |
| Scale axes title | Заголовок оси значений |
| Scaling | Шкалирование |
| Scaling model | Модель шкалирования |
| Scatter plot | Диаграмма рассеивания |
| Scheffe test | Критерий Шеффе |
| Scree plot | График собственных значений |
| Selection | Подбор |
| Sign | Знак |
| Sign test | Критерий знаков |
| Significance | Значимость |
| Significance level | Уровень значимости |
| Similarity | Сходство |
| Skewness | Асимметрия |
| Solution | Решение |
| Source | Источник |
| Spearman's rho | Коэффициент корреляции Спирмена |
| Split-half reliability | Надежность половинного расщепления |
| Square asymmetric matrix | Квадратная асимметричная матрица |
| Square symmetric matrix | Квадратная симметричная матрица |
| Squared Euclidean distance | Квадрат Евклидова расстояния |
| Squared multiple correlation | Квадрат множественной корреляции |
| S-stress | s-стресс |
| S-stress convergence | Величина сходимости s-стресса |
| Standard (std.) error | Стандартная ошибка |
| Standard deviation (std. dev.) | Стандартное отклонение |
| Standard error kurtosis | Стандартная ошибка асимметрии |
| Standard error mean | Стандартная ошибка среднего |
| Standard error skewness | Стандартная ошибка эксцесса |
| Statistical Package for the Social Science (SPSS) | Статистический пакет для социальных наук |
| Statistical test. | Статистический критерий |
| Stepwise method | Пошаговый метод |
| Stimulus | Стимул |
| Stress | Стресс |

| Термин | Перевод |
|---|---|
| Subset | Подмножество |
| Subtitle | Подзаголовок |
| Sum | Сумма |
| Sum of Squares (SS) | Сумма квадратов |
| System missing value | Физически пропущенное значение |
| Territorial map | Карта территории |
| Test | Статистический критерий |
| Test value | Заданное (тестовое) значение |
| Tick | Риска |
| Tied Ranks | Связанные (повторяющиеся) ранги |
| Title | Заголовок |
| Tolerance | Толерантность |
| Total | Итог |
| Tree diagram | Древовидная диаграмма |
| True value | Истинное значение |
| Tukey's test of additivity | Критерий аддитивности Тьюки |
| *t*-value | t-критерий |
| Two-tailed significance | Двусторонняя значимость |
| Unequal-length Spearman—Brown | Критерий неэквивалентных форм Спирмена—Брауна |
| Univariate procedure (approach) | Одномерный метод (подход). |
| Unstandardized coefficient | Нестандартизированный коэффициент |
| User missing value | Логически пропущенное значение |
| Value | Значение |
| Value label | Метка значений |
| Variable (var) | Переменная |
| Variance | Дисперсия |
| Wald test | Критерий Вальда |
| Weight | Вес |
| Weirdness | Предсказуемость |
| Wilcoxon test | Критерий Вилкоксона |
| Wilks' Lambda | Лямбда Уилкса |
| Within-groups linkage | Внутригрупповое связывание |
| Working directory | **Рабочий** каталог |
| Z-score | z-значение (стандартизованное значение) |

# Литература


1. *Айвазян С. А.* Прикладная статистика. Классификация и снижение размерности. М., 1989.
2. *Боровиков В.* Statistica. Искусство анализа данных на компьютере: Для профессионалов. СПб., 2003.
3. *Гмурман В. Е.* Теория вероятностей и математическая статистика. М., 1997.
4. *Гублер Е. В., Генкин А. А.* Применение непараметрических критериев статистики в медико-биологических исследованиях. Л., 1973.
5. *Гусев А. Н.* Дисперсионный анализ в экспериментальной психологии: Учебное пособие для студентов факультетов психологии. М., 2000.
6. *Дэйвисон М.* Многомерное шкалирование: методы наглядного представления данных. М., 1988.
7. *Закс Л.* Статистическое оценивание. М., 1976.
8. *Иберла К.* Факторный анализ. М., 1980.
9. *Кендалл М., Стьюарт А.* Статистические выводы и связи. М., 1973.
10. *Митина О. В., Михайловская И. Б.* Факторный анализ для психологов. М., 2001.
11. *Наследов А. Д.* Математические методы психологического исследования. Анализ и интерпретация данных. СПб., 2004.
12. *Рунион Р.* Справочник по непараметрической статистике. М., 1982.
13. *Сидоренко Е. В.* Методы математической обработки в психологии. СПб., 1996.
14. Справочник по прикладной статистике. В 2-х т. / Под ред. *Э. Ллойда, У. Ледермана, Ю. Тюрина.* М., 1989.
15. *Суходольский Г. В.* Основы математической статистики для психологов. СПб., 1998.
16. *Суходольский Г.* В. Математические методы психологии. СПб., 2003.
17. *Терехина А. Ю.* Анализ данных методами многомерного шкалирования. М., 1986.
18. Факторный, дискриминантный и кластерный анализ / *Дж.-О. Ким, Ч.* У. *Мьюллер,* У. *Р. Клекка* и др. М., 1989.
19. *Харман Г.* Современный факторный анализ. М., 1972.
20. *Agresti, Alan* (1996). An Introduction to Categorical Data Analysis. New York: Wilely.
21. *Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W.* (1975). Discrete multivariate analysis. Cambridge, MA: MIT Press.

# Алфавитный указатель

## Л



## М



## Н



## О



**Наследов Андрей Дмитриевич** — кандидат психологических наук, доцент факультета психологии СПбГУ; специалист в области методов анализа данных психологических исследований. В книге отражен многолетний научный и педагогический опыт автора в применении компьютера для анализа и интерпретации данных.

Эта книга представляет собой уникальный самоучитель по анализу данных психологических и социальных исследований в программе SPSS. Прочитав ее, вы сможете овладеть всеми основными методами анализа, даже не будучи специалистом в области информационных технологий. Основное содержание книги составляют пошаговые инструкции по реализации различных видов математико-статистического анализа в SPSS. Особое внимание уделяется получаемым результатам и их интерпретации. Каждый шаг анализа и получаемый результат иллюстрируются конкретными примерами с использованием файлов, которые читатель может найти на веб-сайте издательства по адресу http://www.piter.com/download. Вот лишь некоторые темы, освещаемые в этой книге:

- первые шаги: ввод в компьютер и правильная организация исходных данных;
- оформление результатов анализа в виде таблиц и графиков;
- простейшие методы первичного анализа данных;
- параметрические и непараметрические методы проверки гипотез;
- все варианты дисперсионного анализа;
- анализ надежности;
- регрессионный, факторный и дискриминантный анализ;
- многомерное шкалирование;
- простая, нелинейная, множественная и логистическая регрессии;
- логлинейный анализ многомерных таблиц сопряженности.







**Заказ книг:**

197198, Санкт-Петербург, а/я 619; тел.: (812)103-73-74, postbook@piter.com

61093, Харьков-93, а/я 9130; тел.: (057)712-27-05, piter@tender.kharkov.ua

**www.piter.com** — вся информация о книгах и веб-магазин